Semi-parametric Bivariate Polychotomous Ordinal Regression

Francesco Donat* and Giampiero Marra

Department of Statistical Science
University College London

October 21, 2015

Abstract

A pair of polychotomous random variables $(Y_1, Y_2)^\top =: \boldsymbol{Y}$, where each Y_j has a totally ordered support, is studied within a penalized Generalized Linear Model framework. We deal with a triangular generating process for $\boldsymbol{Y}$, a structure that has been employed in the literature to control for the presence of residual confounding. Differently from previous works, however, the proposed model allows for a semi-parametric estimation of the covariate-response relationships. In this way, the risk of model mis-specification stemming from the imposition of fixed-order polynomial functional forms is also reduced. The proposed estimation methods and related inferential results are finally applied to study the effect of education on alcohol consumption among young adults in the UK.

Key-words: Alcohol (mis)use; Bivariate Systems of Equations; Ordinal Responses; Penalized GLM; Regression Splines.

1 Introduction

Polychotomous ordinal data arise in many areas of statistical analysis and are particularly frequent in surveys and observational studies. Several questions may be asked to measure people's feelings on a matter of interest, as well as some relevant information reported on a monotonic scale. Examples include individuals' perceived social class or their educational attainments. Since it is usually acknowledged that these types of data possess levels that can be "naturally" ordered, it is desirable to account for this feature in the model's representation and estimation. Specific methodologies were developed to address this issue, starting from the seminal works of Aitchison and Silvey (1957) and Snell (1964), up to their modern forms of the Cumulative Link Models (CLM; McCullagh, 1980) in which ordinal responses are expressed within the wider class of Generalized Linear Models (GLMs; Nelder and Wedderburn, 1972). An interesting historical review discussing merits (and limits) of each of the above contributions can be found in the monograph of Greene and Hensher (2010).

This paper deals with a bivariate system of polychotomous outcomes, $\boldsymbol{Y} := (Y_1, Y_2)^\top$, where each Y_j, $j = 1, 2$, is measured on the ordinal scale. To fix ideas, and recalling that many discrete data can be modelled as a coarse version of a continuous latent random variable Y_j^* (e.g. McKelvey and Zavoina, 1975, Anderson and Philips, 1981), we anticipate that the aim of the article is to estimate and to make inference from a model with the following structure

$$\begin{array}{rcl} Y_1^* & = & \mathbf{x}_1^\top \boldsymbol{\beta}_1 + s_{1,1}(\mathrm{v}_{1,1}) + \cdots + s_{1,L_1}(\mathrm{v}_{1,L_1}) + \epsilon_1 \\ Y_2^* & = & \psi Y_1^* + \mathbf{x}_2^\top \boldsymbol{\beta}_2 + s_{2,1}(\mathrm{v}_{2,1}) + \cdots + s_{2,L_2}(\mathrm{v}_{2,L_2}) + \epsilon_2 \end{array}, \tag{1}$$

*Corresponding Author. University College London, Gower Street, London, WC1E 6BT, UK. Tel.: +44 (0)2076791223. E-mail: f.w.donat@gmail.com

where the s_{j,l_j} are unknown smooth functions appropriately represented and fitted and $\psi \in \mathbb{R}$. Upon setting $\boldsymbol{Y}^* := (Y_1^*, Y_2^*)^\top$, $\mathbf{X} := \text{diag}(\mathbf{x}_1^\top, \mathbf{x}_2^\top)$, $\boldsymbol{\beta} := \text{vec}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $\boldsymbol{\epsilon} := (\epsilon_1, \epsilon_2)^\top$ we re-write (1) in the more compact form $\boldsymbol{\Gamma}\boldsymbol{Y}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ or

$$\mathbf{L}\boldsymbol{Y}^* = \mathbf{L}\boldsymbol{\Gamma}^{-1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \qquad \boldsymbol{\epsilon} \sim \mathcal{N}_2(\mathbf{0}_2, \boldsymbol{\Omega}), \tag{2}$$

where

$$\boldsymbol{\Omega} := \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \boldsymbol{\Gamma} := \begin{bmatrix} 1 & 0 \\ -\psi & 1 \end{bmatrix}, \quad \mathbf{L} := \begin{bmatrix} 1 & 0 \\ 0 & (\sqrt{1 + 2\psi\rho + \psi^2})^{-1} \end{bmatrix} \quad \text{and}$$

$\rho \in [-1, 1]$ is the correlation coefficient. It follows that $\mathbf{L}\boldsymbol{Y}^* \sim \mathcal{N}_2(\mathbf{L}\boldsymbol{\Gamma}^{-1}\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \mathbf{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{-\top}\mathbf{L}^\top$. Model (2) represents the so-called reduced-form of system (1) which is termed triangular (or recursive) given that $\boldsymbol{\Gamma}$ is a lower triangular matrix.

Apart from a pure methodological interest, the study of (2) is motivated by the practical issue of analysing data affected by residual confounding. This is a situation where an unknown or not readily quantifiable variable, or set of variables, is associated with both an ordinal response of interest and an ordinal treatment. When not adequately controlled for, unmeasured confounders may pose serious limitations to the use of standard estimators as they usually yield inconsistent estimates. An analogous bivariate system of equations for dichotomous outcomes addressing this problem has been recently discussed by Radice et al. (2015). At present, the only alternatives available to model ordinal polychotomous responses in a similar (albeit purely parametric) fashion comprise the routines of Sajaia (2008) and the mixed effects version proposed by Buscha and Conte (2014), both for the `STATA` computational environment (StataCorp, 2015). A first contribution of our paper, therefore, concerns the development of an approach for fitting system (1), which permits for a semi-parametric estimation of the covariate-response relationships. This allows us to determine the functional form of covariate effects from the data without the imposition of finite-order polynomials, hence reducing the risk of mis-specification. Moreover, semi-parametric modelling avoids categorising continuous variables into groups based on intervals or frequencies. This approach, which is often employed in empirical studies, is not free from disadvantages; it introduces the issue of defining cut-points, and assumes *a priori* that the relationship between the response and the categorised covariates is flat within the chosen intervals (Royston and Altman, 1994).

In principle, once a distributional assumption for the latent random vector $\boldsymbol{Y}^*$ is made, and an observational rule for the manifest polychotomous responses established, the likelihood function of the model can be easily set up and the parameters estimated. The approach we take here, however, is slightly different and more general. In line with Peyhardi et al. (2014), we specify a GLM class for bivariate ordinal responses defined by the triplet $(r, F_2, \mathbf{Z})$, where F_2 and $\mathbf{Z}$ are a 2-variate distribution function and the design matrix, respectively, and r is a map characterising the types of response vector. We then describe the class of penalized GLMs and show that (2) can be specified as an instance of it. In this way, a generic algorithm for the estimation and inference of any penalized GLMs endowed with the $(r, F_2, \mathbf{Z})$ representation can be developed, and hence potentially applied to any other multivariate model for discrete responses with semi-parametric covariate effects. At a smaller scale, this is already achieved in this paper: we discuss the representations corresponding to a mixture of dichotomous and polychotomous outcomes, as well as some other models nested in the triangular structure. For instance, our framework also comprises the seemingly unrelated regression equations model (SURE) of Hillmann et al. (2014), which is recovered by setting $\mathbf{L} = \mathbf{I}_2$, and allows for the estimation of a bivariate system of independent ordinal probit regressions.

After having represented the triangular structure in a suitable penalized GLM form, Section 3 is devoted to the description of the corresponding estimation algorithm. It is worth stressing that the triplet $(r, F_2, \mathbf{Z})$ is all it is needed for this scope, since it already incorporates the information concerning the model specification, link function used, and types of responses.

In this way, the description of a more general model will enable us to develop an algorithm suitable for any other model belonging to the class. The approach we follow is analogous to that of Vector Generalized Additive Models (VGAM; Yee and Wild, 1996), and that of structured additive regressions models by Klein et al. (2015) and Klein and Kneib (2015). All the necessary computational routines are incorporated in the `R` function `SemiParCLM`, which is available from the website of this article. Finally, our model is illustrated in Section 5 using data from the BCS70 dataset (UCL Institute of Education. Centre for Longitudinal Studies, 2007). The aim of this study is to quantify the effect of education on alcohol consumption among young adults in the UK.

2 A GLM Representation for Bivariate Ordinal Responses

Let us assume that we observe realisations from the distribution of a bivariate random vector $\boldsymbol{Y} = (Y_1, Y_2)^\top$ with discrete support $\mathcal{K} := \mathcal{K}_1 \times \mathcal{K}_2$, such that $(\mathcal{K}_j, \preceq)$ is totally ordered for any $j \in \mathcal{J} = \{1, 2\}$. Specifically, we consider the set $\mathcal{K}_j := \{1, \ldots, k_j, \ldots, K_j\}$ with $\#(\mathcal{K}_j) = K_j < \infty$, where k_j represents a natural number. We then say that variable Y_j shows finite K_j levels. Notice that the totality assumption implies the comparability of each k_j with respect to all the remaining elements in $\mathcal{K}_j \setminus \{k_j\}$. In other words, the proposed methodology is only applicable in those situations where it is possible to state whether $\bar{k}_j \preceq k_j$ or $k_j \preceq \bar{k}_j$ for any $k_j, \bar{k}_j \in \mathcal{K}_j$. For example, this may not be the case in surveys foreseeing the possibility to tick the "don't know" box. Whenever this instance is likely to occur, more appropriate models for partially ordered responses have to be employed, like the one discussed by Zhang and Ip (2012).

Covariate information is collected in the vector $\boldsymbol{x} := \text{vec}(\boldsymbol{x}_1, \boldsymbol{x}_2)$, where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are the available regressors. It is then licit to set up a model relating the conditional probability $\pi_k := \mathbb{P}[\boldsymbol{Y} = k | \boldsymbol{X} = \boldsymbol{x}]$, with $k := (k_1, k_2)^\top \in \mathcal{K}$, to $\boldsymbol{x}$ through the GLM form (Peyhardi et al., 2014)

$$\boldsymbol{\pi} = \boldsymbol{g}^{-1}(\boldsymbol{\eta}) := (\boldsymbol{r}^{-1} \circ \mathcal{F})(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{K-1}) \in [0, 1]^{\#(\mathcal{K})-1}, \tag{3}$$

where $\mathcal{F}(\boldsymbol{\eta}) := (F_2(\boldsymbol{\eta}_1), \ldots, F_2(\boldsymbol{\eta}_{K-1}))^\top$, $F_2 : \mathbb{R}^2 \longrightarrow (0, 1)$ denotes any fully-specified bivariate distribution function and $K := (K_1, K_2)^\top$. A bivariate CLM for polychotomous ordinal responses is then recovered by setting $\boldsymbol{r}(\boldsymbol{\pi}) := (r(\pi_k))_{k \in \mathcal{K} \setminus \{K\}}$ where, for each k,

$$r(\pi_k) := \mathbb{P}[Y_1 \preceq k_1, Y_2 \preceq k_2 | \boldsymbol{X} = \boldsymbol{x}] = \sum_{\tilde{k}_1 \preceq k_1} \sum_{\tilde{k}_2 \preceq k_2} \pi_{\tilde{k}_1, \tilde{k}_2}.$$

The array $\boldsymbol{\eta}_k := (\eta_{1,k_1}, \eta_{2,k_2})^\top \in \mathbb{R}^2$ defines the linear predictor of the model, and embodies the functional form of the covariate effects. Since this is pivotal in our proceeding discussion, it will be analysed more precisely in Section 2.1. In line with the multivariate nature of the model, the generic pair $(k_1, k_2) \in \mathcal{K}$ is assumed to follow a lexicographical order, that is $(\bar{k}_1, \bar{k}_2) \preceq (k_1, k_2)$ if and only if $\bar{k}_1 \preceq k_1$ or $(\bar{k}_1 = k_1 \ \wedge \ \bar{k}_2 \preceq k_2)$ for $\bar{k}_j, k_j \in \mathcal{K}_j$. We stress that:

Remark 1. *Any regression model for ordinal outcomes sets two constraints in representation (3). The obvious one requires $r(\pi_k) = r(\pi_{\bar{k}}) + \pi_k \geq r(\pi_{\bar{k}})$ for $\bar{k} := (k_1, k_2 - 1) \preceq (k_1, k_2) =: k$; whereas $\boldsymbol{\eta}_{\bar{k}} \leq \boldsymbol{\eta}_k$ is needed for all $\bar{k} \preceq k$ and $\bar{k}, k \in \mathcal{K}$. In particular, the latter can be thought of as a model coherency condition and is introduced to ensure that the order relations implied by the set $\mathcal{K}$ are maintained in the domain of the linear predictor, $\mathbb{R}^2$.*

To meet these requirements, let us set a pair of cut points (or threshold parameters), collected in the vector $\boldsymbol{c}_k$ of $\{(c_{1,k_1}, c_{2,k_2})^\top \in \mathbb{R}^2 | c_{j,\bar{k}_j} \leq c_{j,k_j}, \forall \bar{k}_j \preceq k_j, k_j \in \mathcal{K}_j, \forall j\}$, and such that $c_{j,K_j} = \infty$ and $c_{j,0} := c_{j,1-1} = -\infty$. We consequently define a bivariate probit regression for ordinal responses as

$$r(\pi_k) = \Phi_2(\boldsymbol{c}_k - \mathbf{X}\boldsymbol{\beta}) = \Phi_2(\mathbf{Z}\boldsymbol{\beta}_k), \tag{4}$$

where $\mathbf{Z} := \text{diag}(\mathbf{z}_1^\top, \mathbf{z}_2^\top)$ is the analogue of the design matrix, $\mathbf{z}_j := (1, -\mathrm{x}_{j,1}, \ldots, -\mathrm{x}_{j,M_j})$, and $\boldsymbol{\beta}_k := \text{vec}(\boldsymbol{\beta}_{1,k_1}, \boldsymbol{\beta}_{2,k_2})$, $\boldsymbol{\beta}_{j,k_j} := (c_{j,k_j}, \beta_{j,1}, \ldots, \beta_{j,M_j})^\top \in \mathbb{R}^{M_j+1}$, is the vector of regression coefficients. M_j is used to denote the number of covariates included in equation j. Finally, the linear predictors are given by $\boldsymbol{\eta}_k := \mathbf{Z}\boldsymbol{\beta}_k$, so that GLM form (4) can be characterised by the triplet $(r, F_2, \mathbf{Z})$. Notice that we have set $F_2 \equiv \Phi_2$ in the proposed model specification.

The above definition of the cut points relies on the weak monotonicity assumption of $\{c_{j,k_j}\}_{k_j}$ for all j. In fact, although Dale (1986) required this sequence to be *strictly* increasing to ensure that each π_k is positive, we regard this condition too stringent, as it eventually adds a further unnecessary constraint to the likelihood function. Admittedly, as Haberman (1980) pointed out for the univariate case, wherever two subsequent cut points are congruent (e.g., when zero counts are observed for a given level) the Maximum Likelihood Estimator (MLE) is located at the boundary of the parameter space. Notwithstanding, since the estimates so obtained are still admissible *per se*, it seems to us that the exclusion of the case $c_{j,k_j} = c_{j,k_j+1}$ is formally restrictive and thus to be avoided. Some alternative estimators to the MLE dealing with this issue have been recently proposed by Kosmidis (2014) for univariate CLMs.

Remark 2. *Although our focus is on the modelling of ordinal responses, our methodology is immediately applicable also to mixtures of dichotomous and polychotomous variables. To see this, let us first decompose $r = r_2 \circ r_1$, where the subscripts correspond to the elements of the 2-dimensional vector $\boldsymbol{Y}$ they refer to. Moreover, the inclusion of a binary outcome in (3), say $Y_{\bar{\jmath}}$, corresponds to define $r_{\bar{\jmath}}$ as $\pi_{k_{\bar{\jmath}}} \mapsto \pi_{k_{\bar{\jmath}}}$, the identity map. Then it follows*

$$(r_{\bar{\jmath}} \circ r_j)(\pi_k) = \pi_{k_{\bar{\jmath}},1} + \cdots + \pi_{k_{\bar{\jmath}},k_j}.$$

Notice that the fact we have put $k_{\bar{\jmath}}$ before k_j was just for notational convenience. In fact, it is indifferent the order in which the different types of variables appear. More formally, since $r_{\bar{\jmath}}$ is the identity map, we have that the function composition is commutative:

$$(r_j \circ r_{\bar{\jmath}})(\pi_k) = r_j(\pi_k) = r_{\bar{\jmath}}(r_j(\pi_k)) = (r_{\bar{\jmath}} \circ r_j)(\pi_k)$$

for $j, \bar{\jmath} \in \mathcal{J}$ and every $k \in \mathcal{K}$.

In the proceeding discussion, we extend representation (4) to account for semi-parametric model components, and develop a generic estimation algorithm for a bivariate system of polychotomous ordinal responses expressible in the $(r, F_2, \mathbf{Z})$ form.

2.1 Regression Spline Representation

Each linear predictor $\boldsymbol{\eta}_k$ can be specified to embody different types of covariate effects. In this work, additive non-parametric effects of the continuous regressors v_{j,l_j} are represented using regression splines (Eilers and Marx, 1996). Let us assume we observe a sample of n individuals indexed by the subscript i, and let $\{\mathrm{v}_{j,l_j,(1)}, \ldots, \mathrm{v}_{j,l_j,(n)}\}$ be the ordered vector of corresponding observations. Thus, provided that we can choose a rich enough set of basis functions, $\boldsymbol{b}_{j,l_j}$, delimited by $H_j + 1$ knot points in the interior of $[\mathrm{v}_{j,l_j,(1)}, \mathrm{v}_{j,l_j,(n)}]$, we approximate

$$s_{j,l_j}(\mathrm{v}_{j,l_j,i}) \approx \boldsymbol{\delta}_{j,l_j}^\top \boldsymbol{b}_{j,l_j}(\mathrm{v}_{j,l_j,i}) \in \mathbb{R}.$$

Specifically, $s_{j,l_j} : \mathbb{R} \longrightarrow \mathbb{R}$ is restricted to be a smooth function, $\boldsymbol{b}_{j,l_j}(\mathrm{v}_{j,l_j,i}) := (b_{j,l_j,h_j}(\mathrm{v}_{j,l_j,i})) \in \mathbb{R}^{H_j}$, and $\boldsymbol{\delta}_{j,l_j} \in \mathbb{R}^{H_j}$ is a parameter vector associated to s_{j,l_j}. Basis functions are usually chosen to have convenient mathematical properties and good numerical stability. Among the various functions supported by our implementation, the B-splines, cubic regression and thin-plate regression splines are the most widely used in applications (e.g. Ruppert et al., 2003; Wood, 2003). To achieve functions' identification, the centering constraint $\mathbf{1}_n^\top \boldsymbol{s}_{j,l_j} = 0$ is imposed, where $\boldsymbol{s}_{j,l_j}$ denotes the vector whose i-th element is $s_{j,l_j}(\mathrm{v}_{j,l_j,i})$. This approach is incorporated

automatically in our model estimation through the parsimonious method outlined in Wood (2006).

To recover a more compact and comprehensive representation of the linear predictors, we set $\boldsymbol{\beta}_{[j,l_j]} := \boldsymbol{\delta}_{j,l_j}$ which represents the sub-vector of $\boldsymbol{\beta}_j$ referring to the (j,l_j)-th smooth and, accordingly, $\mathbf{X}_{[j,l_j]} \in \mathbb{R}^{n \times H_j}$ is the matrix whose i-th row is given by $\boldsymbol{b}_{j,l_j}^{\top}(\mathrm{v}_{j,l_j,i})$. Then, we can write the linear predictor of the j-th response as

$$\boldsymbol{\eta}_j = \boldsymbol{c}_j - \mathbf{X}_{j,1}\boldsymbol{\beta}_{j,1} - \cdots - \mathbf{X}_{j,M_j}\boldsymbol{\beta}_{j,M_j} = \mathbf{Z}_j\boldsymbol{\beta}_j \in \mathbb{R}^n,$$

where $\mathbf{Z}_j := (\mathbf{1}_n, -\mathbf{X}_{j,1}, \ldots, -\mathbf{X}_{j,m_j}, \ldots, -\mathbf{X}_{j,M_j})$, $\boldsymbol{\beta}_j := \mathrm{vec}(\boldsymbol{c}_j, \boldsymbol{\beta}_{j,1}, \ldots, \boldsymbol{\beta}_{j,M_j})$ and $\boldsymbol{c}_j := (c_{j,k_j,i})_i \in \mathbb{R}^n$. So re-stated, the linear predictors can be employed to incorporate both non- and purely parametric covariate effects. A modelling approach of this kind is termed semi-parametric in the statistical literature.

2.2 The Triangular Ordered Probit Model

The previous sections have described a generic model for a bivariate ordinal polychotomous random vector. In what follows, we qualify the structure of the triangular model of interest.

Motivation Residual confounding is a relatively frequent issue in observational studies. It occurs whenever the association between a response and one (or more) of its relevant regressor(s) is distorted by the presence of an unobserved third variable which affects simultaneously the two. Such covariates are termed endogenous in the econometric literature. A researcher would be particularly interested in controlling for pertinent unmeasured confounders as they usually lead to inconsistent estimates for the whole parameter vector. In experimental studies, one possible solution is the assignment of the relevant treatment to individuals via a randomisation mechanism, whose functioning is independent of any other factor (e.g. Frosini, 2006). However, this may not be feasible in situations where the experiment design would raise ethical or legal issues, as it is frequently the case in observational studies. Models dealing with this problem have been proposed in the literature. Cox and Wermuth (2004) and Wermuth and Cox (2008), for example, described the direct confounding effect by means of graphical models. In this setting, they quantified the distortion from endogenous covariate effects under the regular assumptions of continuous responses and a generating process represented by a triangular system of equations.

In line with the bivariate recursive model introduced by Heckman (1978) for binary outcomes, we consider the instance of an endogenous variable Y_1 that is assumed to have an impact on the response of interest Y_2. Each of them is defined on the discrete and totally ordered support $\mathcal{K}_j$, with $\#(\mathcal{K}_j) \geq 2$. In the empirical study, we argue that individuals' education attainments are potentially endogenous in explaining their weekly alcohol intake, because both affected by a common subjective attitude. This underlying variable is recognised to be time preference in the relevant economic literature. A bivariate system of equations is then employed to describe this situation. The corresponding generating process – expressed in terms of the latent variable formulation – is the one previously given in (1) and (2). Notice that, in addition to the usual distributional assumption (e.g. Greene and Hensher, 2010), a further condition in the form of an exclusion restriction has to be imposed in the model to achieve identification (Sajaia, 2008, Buscha and Conte, 2014). This allows us to qualify the dependence of Y_1 with a relevant variable which is independent of (i) $Y_2|Y_1$, and (ii) the unmeasured confounder. We argue, for example, that the British Ability Scale score possesses these characteristics in the real data illustration of Section 4.

Model Representation As most models for discrete data, ordinal polychotomous variables can also be motivated by means of a generating latent and continuous random vector, $\boldsymbol{Y}^*$, with

Model	$r(\pi_k)$	$F_2(\boldsymbol{\eta}_k)$	$\boldsymbol{\eta}_k(\mathbf{Z})$
Triangular	$\sum_{\tilde{k}_1 \leq k_1} \sum_{\tilde{k}_2 \leq k_2} \pi_{\tilde{k}_1, \tilde{k}_2}$	$\Phi_2(\eta_{1,k_1}, \eta_{2,k_2}; \boldsymbol{\Sigma})$	$\mathbf{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}_k$
SURE	$\sum_{\tilde{k}_1 \leq k_1} \sum_{\tilde{k}_2 \leq k_2} \pi_{\tilde{k}_1, \tilde{k}_2}$	$\Phi_2(\eta_{1,k_1}, \eta_{2,k_2}; \boldsymbol{\Omega})$	$(\eta_{1,k_1}, \eta_{2,k_2})^\top$
Independent	$\sum_{\tilde{k}_1 \leq k_1} \sum_{\tilde{k}_2 \leq k_2} \pi_{\tilde{k}_1, \tilde{k}_2}$	$\Phi(\eta_{1,k_1})\Phi(\eta_{2,k_2})$	$(\eta_{1,k_1}, \eta_{2,k_2})^\top$
$\mathcal{K}_1 = \{0, 1\}$	$\pi_{k_1,1} + \cdots + \pi_{k_1,k_2}$	$\Phi_2(\eta_{1,k_1}, \eta_{2,k_2}; \boldsymbol{\Sigma})$	$\mathbf{L}\boldsymbol{\Gamma}^{-1}((-1)^{\mathbb{1}_{k_1=0}}\eta_{1,k_1}, \eta_{2,k_2})^\top$
$\mathcal{K}_2 = \{0, 1\}$	$\pi_{1,k_2} + \cdots + \pi_{k_1,k_2}$	$\Phi_2(\eta_{1,k_1}, \eta_{2,k_2}; \boldsymbol{\Sigma})$	$\mathbf{L}\boldsymbol{\Gamma}^{-1}(\eta_{1,k_1}, (-1)^{\mathbb{1}_{k_2=0}}\eta_{2,k_2})^\top$

Table 1: $(r, F_2, \mathbf{Z})$ characterisation corresponding to structure (5) under different model specifications. The SUR equations set $\psi = 0$, hence $\boldsymbol{\Gamma} = \mathbf{L} = \mathbf{I}_2$ and $\boldsymbol{\Sigma} := \mathbf{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{-\top}\mathbf{L}^\top = \boldsymbol{\Omega}$. Two independent ordinal probit models are recovered by letting $\psi = \rho = 0$ so that $\boldsymbol{\Sigma} = \mathbf{I}_2$. The last two rows report the representation corresponding to mixtures of dichotomous and polychotomous responses in the triangular model as stated in Remark 2. Notice that, since only $K_j - 1$ cut points are effectively estimated, the condition $c_{j,0} := 0$ is usually set for the equation corresponding to the binary response, and the intercept is now estimable. The label $\boldsymbol{\eta}_k \in \mathbb{R}^2$ has been used to denote the i-th row of $\boldsymbol{\eta}$, which in turn depends on the level $k \in \mathcal{K}$.

support the extended real plane, through the equivalence (McKelvey and Zavoina, 1975)

$$\{\boldsymbol{Y} = (k_1, k_2) \subseteq \mathcal{K}\} \iff \{\boldsymbol{Y}^* \in [c_{1,k_1-1}, c_{1,k_1}] \times [c_{2,k_2-1}, c_{2,k_2}] \subseteq \mathbb{R}^2\},$$

where the Cartesian product defines the non-overlapping rectangles in $\mathbb{R}^2$ whose vertices are the cut points. Using (2), and by noticing that $\mathbf{L}$ is positive-definite, since $1 + 2\psi\rho + \psi^2 = (1 - \rho^2) + (\rho + \psi)^2 > 0$ and its determinant positive, it holds that

$$\begin{aligned} \{\boldsymbol{Y} \preceq k\} &\iff \{\boldsymbol{\Gamma}\boldsymbol{Y}^* \leq \boldsymbol{c}_k\} \iff \{\mathbf{L}\boldsymbol{Y}^* \leq \mathbf{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{c}_k\} \\ &\iff \{\mathbf{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{\epsilon} \leq \mathbf{L}\boldsymbol{\Gamma}^{-1}(\boldsymbol{c}_k - \mathbf{X}\boldsymbol{\beta})\}, \end{aligned}$$

where the equivalence is established under coherency. Hence, given the assumed Standard Normal distribution of the stochastic model components, the proposed triangular structure for a sample of size n corresponds to the setting of

$$r(\boldsymbol{\pi}) = \Phi_2\left(\mathbf{Z}\boldsymbol{\beta}(\mathbf{L}\boldsymbol{\Gamma}^{-1})^\top; \boldsymbol{\Sigma}\right) \in [0, 1]^n \qquad \boldsymbol{\Sigma} = \mathbf{I}_n \otimes \mathbf{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{-\top}\mathbf{L}^\top, \tag{5}$$

where $\boldsymbol{\pi} := (\pi_1, \ldots, \pi_n)^\top$, $\pi_i := \mathbb{P}[\mathrm{y}_{1,i} = k_1, \mathrm{y}_{2,i} = k_2 | \mathbf{X}]$, $\mathbf{Z} := (\mathbf{Z}_1 | \mathbf{Z}_2)$ and $\boldsymbol{\beta} := \mathrm{diag}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in \mathbb{R}^{M \times 2}$, with $M := M_1 + M_2$. Notice that the assumed recursive structure implies a predictor of the form $\boldsymbol{\eta} := \mathbf{Z}\boldsymbol{\beta}(\mathbf{L}\boldsymbol{\Gamma}^{-1})^\top \in \mathbb{R}^{n \times 2}$. Furthermore, since the quantity $\mathbf{L}\boldsymbol{\Gamma}^{-1}$ involves a non-linear combination of the elements of the p-dimensional vector $\boldsymbol{\vartheta} := \mathrm{vec}(\boldsymbol{c}_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \rho) \in \mathbb{R}^{p-1} \times [-1, 1]$, it follows that $\boldsymbol{\eta}$ is non-linear in the parameter vector. Therefore, strictly speaking, the term "linear predictor" does not apply for the proposed triangular structure, and one needs to be careful in exploiting the GLM properties of this model. As we can see in the next section, some extra terms in the expressions for the score and Hessian have to be accounted for.

Finally, all the relevant model specifications nested in (5) are summarised in Table 1, in which the corresponding $(r, F_2, \mathbf{Z})$ forms are detailed. Estimation can hence proceed by employing a generic algorithm as detailed in the next section. In particular, the seemingly unrelated regression (SUR) representation is recovered by setting $\psi = 0$. This form is usually employed for joint modelling inter-related outcomes or symmetry in the responses. This is the case, for instance, of the estimation of the injuries sustained by two people in the same car accident (Yamamoto and Shankar, 2004), or the intensity of a certain disease in humans' left and right eyes (Kim, 1995).

3 Estimation Methods and Inference

In this paper the random vector $\boldsymbol{Y}|\boldsymbol{X}$ is assumed to follow a Categorical distribution, which is a member of the exponential family of distributions. Using a random sample of conditionally

independent responses given the regressors, we write the log-likelihood function for generic model (3) as

$$\ell(\boldsymbol{\vartheta}|\mathbf{y}_1,\mathbf{y}_2,\mathbf{X}_1,\mathbf{X}_2)=\sum_{i=1}^{n}\sum_{k\in\mathcal{K}}\mathbb{1}_{\mathrm{y}_{1,i}=k_1}\mathbb{1}_{\mathrm{y}_{2,i}=k_2}\log\pi_k(\mathbf{x}_{1,i},\mathbf{x}_{2,i}),$$

where $\mathbf{x}_{j,i}^\top$ is the i-th row of matrix $\mathbf{X}_j$. Notice that the above expression remains valid irrespectively of the model actually used. In fact, it is the computation of each π_k that depends on the specific $(r,F_2,\mathbf{Z})$ form (these are all detailed in Table 1). For a bivariate polychotomous ordinal regression we have

$$\pi_k=r^{-1}(r(\pi_k))=r(\pi_{k_1-1,k_2-1})-r(\pi_{k_1-1,k_2})-r(\pi_{k_1,k_2-1})+r(\pi_{k_1,k_2}),$$

where each addendum can be computed as an instance of (5) for the triangular model. For every i, we set

$$\begin{aligned}\boldsymbol{\eta}_k &:= (\eta_{k_1-1,k_2-1},\eta_{k_1-1,k_2},\eta_{k_1,k_2-1},\eta_{k_1,k_2},\rho)^\top\in\mathbb{R}^4\times[-1,1],\\ \boldsymbol{r}_k &:= (r(\pi_{k_1-1,k_2-1}),r(\pi_{k_1-1,k_2}),r(\pi_{k_1,k_2-1}),r(\pi_{k_1,k_2}))^\top\in\mathbb{R}^4,\end{aligned}$$

so that the analytical expressions for score and Hessian are computed as

$$\nabla_{\boldsymbol{\vartheta}}\ell_i(\boldsymbol{\vartheta})=\frac{\partial\boldsymbol{\eta}_k}{\partial\boldsymbol{\vartheta}}\left[\frac{1}{\pi_k}\frac{\partial\mathcal{F}_k}{\partial\boldsymbol{\eta}_k}\frac{\partial\pi_k}{\partial\boldsymbol{r}_k}\right]=\mathbf{D}_i^\top\mathbf{u}_i=:\boldsymbol{g}_i \tag{6}$$

and

$$\nabla_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}^\top}\ell_i(\boldsymbol{\vartheta})=\mathbf{D}_i^\top\left[\frac{1}{\pi_k}\frac{\partial^2\mathcal{F}_k}{\partial\boldsymbol{\eta}_k\partial\boldsymbol{\eta}_k^\top}-\mathbf{u}_i\mathbf{u}_i^\top\right]\mathbf{D}_i+\frac{\partial^2\boldsymbol{\eta}_k}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\top}\mathbf{u}_i=\mathbf{D}_i^\top\mathbf{W}_i\mathbf{D}_i+\mathbf{K}_i. \tag{7}$$

Further, notice that wherever *linear* predictors are used in the model (i.e. $\psi=0$), $\mathbf{K}_i$ is structurally equal to $\mathbf{0}_p$, and $\mathbf{D}_i$ reduces to the usual design matrix. By appropriately extending the approach of Yee and Wild (1996), we define the arrays $\mathbf{W}:=-\mathrm{diag}(\mathbf{W}_1,\ldots,\mathbf{W}_n,\mathbf{K})$, $\mathbf{D}:=(\mathbf{D}_1^\top|\cdots|\mathbf{D}_n^\top|\mathbf{I}_p)^\top$ and $\mathbf{u}:=\mathrm{vec}(\mathbf{u}_1,\ldots,\mathbf{u}_n,\mathbf{0}_p)$, where $\mathbf{K}:=\sum_i\mathbf{K}_i$. These quantities are conveniently constructed to give global expressions for the score and Hessian analogous to those of univariate GLMs.

Our model specification also imposes two constraints on the parameters. Correlation ρ is by definition bounded in the closed interval $[-1,1]$, whereas the threshold parameters are restricted to be a monotonic series under Remark 1. To this end, we make use of some transformations commonly employed in the literature. Specifically, ρ is set to its inverse hyperbolic tangent, namely $\tilde{\rho}:=\tanh^{-1}(\rho)\in\mathbb{R}$, whilst the cut points are defined via a squared polynomial as in Sajaia (2008). That is, we have $\tilde{c}_{j,1}=c_{j,1}$ and $\tilde{c}_{j,k_j}:=\sqrt{c_{j,k_j}-c_{j,k_j-1}}$ for any $k_j\in\mathcal{K}_j\setminus\{1\}$ and all j, resulting in $c_{j,k_j}=c_{j,k_j-1}+\tilde{c}_{j,k_j}^2\geq c_{j,k_j-1}$. In line with the discussion of Section 2, we are effectively allowing two subsequent cut points to be congruent wherever $\tilde{c}_{j,k_j}$ is estimated as 0. To avoid clutter in the notation, in what proceeds we do not distinguish between the parameter vector $\boldsymbol{\vartheta}$ and its transformation $\tilde{\boldsymbol{\vartheta}}\in\mathbb{R}^p$, where the latter includes the quantities $\tilde{c}_{j,k_j}$ and $\tilde{\rho}$. Estimation is nonetheless intended to be over $\tilde{\boldsymbol{\vartheta}}$: that is we seek to maximise $\ell(\tilde{\boldsymbol{\vartheta}}|\cdot)$ with respect to $\boldsymbol{\vartheta}$.

3.1 Penalized GLM Form

Classic MLE is not suitable in semi-parametric regression. In fact, the intuitive optimisation of the model log-likelihood may give rise to over-fitted curves if smoothness is not adequately calibrated. To avoid this issue, we introduce in fitting a ridge-type penalty, namely $\mathcal{P}_{j,m_j}:=\lambda_{j,m_j}\boldsymbol{\beta}_{j,m_j}^\top\overline{\mathbf{S}}_{j,m_j}\boldsymbol{\beta}_{j,m_j}$, whose role is to enforce certain properties of the (j,m_j)-th covariate. The tuning parameters $\lambda_{j,m_j}\in[0,\infty)$ govern the trade-off between smoothness and fit. At one

extreme, $\lambda_{j,m_j} = 0$ assigns no penalty to the regression coefficients $\boldsymbol{\beta}_{j,m_j}$ and the corresponding estimated effect may interpolate the data points. At the other, $\lambda_{j,m_j} \to \infty$ results in the estimation of a straight line, a situation where the smoothness is maximal. The smoothing parameters are thus of paramount importance in any regression spline modelling, and need to be reliably estimated within the system.

The proposed representation is flexible enough to accommodate both purely and non-parametric effects of the (j, m_j)-th covariate, where the former is achieved by setting $\overline{\mathbf{S}}_{j,m_j} = \mathbf{0}$. For non-parametric curve fitting one can specify the symmetric and positive semi-definite penalty matrix as

$$\overline{\mathbf{S}}_{j,m_j} := \int_{V_{j,m_j}} \boldsymbol{b}''_{j,m_j} (\boldsymbol{b}''_{j,m_j})^\top \mathrm{d}v_{j,m_j},$$

a measure of the curvature of the estimated (j, m_j)-th function. Introductions to this roughness penalty approach to curve estimation are given in Green and Silverman (1994) and Wood (2006), to which we refer the reader for details. Finally, after having regularised each penalty matrix to account for the centering constraint of Section 2.1, one can explicitly construct an overall penalisation term for the whole system as $\mathcal{P}_{\boldsymbol{\lambda}} := \boldsymbol{\vartheta}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\vartheta}$, where $\mathbf{S}_{\boldsymbol{\lambda}}$ corresponds to $\overline{\mathbf{S}}_{\boldsymbol{\lambda}}$ padded with zero so that $\boldsymbol{\vartheta}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\vartheta} = \boldsymbol{\beta}^\top \overline{\mathbf{S}}_{\boldsymbol{\lambda}} \boldsymbol{\beta}$, with $\overline{\mathbf{S}}_{\boldsymbol{\lambda}} := \mathrm{diag}(\overline{\mathbf{S}}_{j,m_j})_{m_j,j}$.

3.2 Estimation Given the Smoothing Parameters

Parameter estimation is achieved by alternating two steps in the spirit of the outher itertion algorithm of O'Sullivan et al. (1986). They comprise: (i) the computation of $\boldsymbol{\vartheta}^{[\alpha+1]}$ given any fixed $\boldsymbol{\lambda}^{[\alpha]}$, and (ii) the employment of this estimate to update $\boldsymbol{\lambda}^{[\alpha+1]}$. At convergence, the resulting Maximum Penalized Likelihood Estimator (MPLE) is then

$$\widehat{\boldsymbol{\vartheta}} := \arg\max_{\boldsymbol{\vartheta}} \ell_{\mathrm{p}}(\boldsymbol{\vartheta}, \boldsymbol{\lambda}|\cdot) = \left[\ell(\boldsymbol{\vartheta}|\cdot) - \frac{1}{2} \boldsymbol{\vartheta}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\vartheta} \right]. \tag{8}$$

Notice that the included quadratic form $\mathcal{P}_{\boldsymbol{\lambda}} = \boldsymbol{\vartheta}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\vartheta}$ is positive-semidefinite, and that the unpenalized log-likelihood function does not depend on the smoothing parameters. Hence, joint estimation of $(\boldsymbol{\vartheta}, \boldsymbol{\lambda})$ through the optimisation of (8) would clearly result in over-fitted curves, as the optimal value of $\ell_{\mathrm{p}}(\boldsymbol{\vartheta}, \boldsymbol{\lambda}|\cdot)$ would be reached when $\widehat{\boldsymbol{\lambda}} = \mathbf{0}$.

Although in principle the MPLE can be implemented using any numerical optimisation procedure, works on bivariate discrete response modelling emphasises that considerable gains in precision and computational speed can be achieved by employing a trust-region algorithm (e.g. Marra and Radice, 2013 and Radice et al., 2015). In particular, the $[\alpha]$-th iteration of the routine solves the sub-problem

$$\begin{aligned} \min_{\boldsymbol{p}} \ \widetilde{\ell}_{\mathrm{p}} &= -\left[\ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}) + \boldsymbol{p}^\top \nabla_{\boldsymbol{\vartheta}^{[\alpha]}} \ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}) + \tfrac{1}{2} \boldsymbol{p}^\top \nabla_{\boldsymbol{\vartheta}^{[\alpha]} \boldsymbol{\vartheta}^{[\alpha]\top}} \ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}) \boldsymbol{p} \right] \Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}} \\ &\quad \text{subject to } \|\boldsymbol{p}\| \leq \Delta^{[\alpha]} \\ \boldsymbol{\vartheta}^{[\alpha+1]} &= \boldsymbol{\vartheta}^{[\alpha]} + \boldsymbol{p}^{[\alpha+1]}, \end{aligned} \tag{9}$$

where $\boldsymbol{p}^{[\alpha+1]} := \arg\min_{\boldsymbol{p}} \widetilde{\ell}_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}|\boldsymbol{\lambda}^{[\alpha]}, \cdot)$. The first line of (9) uses a quadratic approximation of the negative log-likelihood about $\boldsymbol{\vartheta}^{[\alpha]}$ (the so-called model function) in order to choose the best step $\boldsymbol{p}^{[\alpha+1]}$ within the ball centered in $\boldsymbol{\vartheta}^{[\alpha]}$ of radius $\Delta^{[\alpha]}$, the trust-region. This step is made particularly precise and quick by using the analytical score and Hessian as computed via (6) and (7).

Trust-region algorithms have a number of advantages for the proposed bivariate system of equations. Recall from Section 2.2 that model estimation requires the imposition of an exclusion restriction to achieve identification. In fact, although it can be argued that identification can

also be achieved by functional form, in this case the log-likelihood may happen to be nearly flat in a non-negligible area around the optimum (e.g. Keane, 1992). This is also the case whenever the excluded covariate is a weak predictor of Y_1. In line-search optimisers, if a given iteration falls in such long plateau regions, the search for a subsequent step, say $\boldsymbol{\vartheta}^{[\alpha+1]}$, can occur far away from the current location $\boldsymbol{\vartheta}^{[\alpha]}$. Nonetheless, the algorithm can still locate that iteration close to $\boldsymbol{\vartheta}^{[\alpha]}$, and only marginal gains in the objective function are obtained. It is also possible that the search happens so far away from $\boldsymbol{\vartheta}^{[\alpha]}$ that the evaluation of (8) is indefinite or not finite. Most algorithms may fail in this case, and user's intervention often required.

Trust-region methods, on the other hand, always solve sub-problem (9) before evaluating the objective function. Specifically, wherever this is not finite at the proposed $\boldsymbol{\vartheta}^{[\alpha+1]}$, the step $\boldsymbol{p}^{[\alpha+1]}$ is rejected, the trust-region shrunken, and the optimisation computed again. The radius is also reduced if there is not agreement between the model and objective functions, that is in case the proposed point in the region is not better than the current one. Reversibly, if such agreement occurs, it is safe to expand the trust region for the next iteration. In summary, $\boldsymbol{\vartheta}^{[\alpha+1]}$ is accepted if it improves on $\boldsymbol{\vartheta}^{[\alpha]}$ and it does not causes problems in the evaluation of $\ell_{\text{p}}(\boldsymbol{\vartheta}^{[\alpha+1]}|\boldsymbol{\lambda}^{[\alpha]})$, whereas the reduction/expansion of $\Delta^{[\alpha+1]}$ is based on the similarity between model and objective functions. This is represented schematically in Algorithm 1. A theoretical background and a general treatment of the algorithm is found in Nocedal and Wright (2006), whereas technical details on the implementation we have followed are given in Geyer (2013). The latter also discusses the necessary modifications to the sub-problem (9) and the radius for ill-scaled variables. It is worthwhile to remark that the discussion in the next section requires some iterations of the optimisation routine to be either of Newton-Raphson or of Fisher scoring-type. Close to the converged solution, the trust-region usually behaves like a classic unconstrained optimisation algorithm (Geyer, 2013; Nocedal and Wright, 2006), and this issue is therefore typically overcome.

Starting values for algorithm initialisation are conveniently fixed at convergence of the corresponding purely parametric version of the model. This practice is efficient and accounts for the presence of unmeasured confounding (which induces parameters' inconsistency), and hence allows us to locate starting values in a region that is reasonably close to the MPLE.

3.3 Smoothness Selection

Once an optimal value for $\ell_{\text{p}}(\boldsymbol{\vartheta}^{[\alpha+1]}|\boldsymbol{\lambda}^{[\alpha]},\cdot)$ has been obtained by the scheme detailed above, we need to employ a proper estimator for $\boldsymbol{\lambda}$. A number of different techniques have been proposed in the literature to estimate smoothing parameters in an automatic way. Among them, the Un-Biased Risk Estimator (UBRE) and the Generalized Cross Validation criterion (GCV, Craven and Wahba, 1979) share a primer position in applied research. In fact, their practical implementation is strengthened by the stable and efficient computational routines introduced by Wood (2004) in the context of GAMs. These have been made applicable and been directly incorporated in our algorithm. In particular, we adapt to the present context the UBRE criterion as the default option for its interpretation in terms of the log-likelihood Akaike Information Criterion (AIC).

Let $\mathbb{R}^{5n} \ni \mathbf{z} := \mathbf{D}\boldsymbol{\vartheta} + \overline{\mathbf{W}}^{-1}\mathbf{u}$ be the pseudo-data vector associated with the un-penalized model, as based on the Fisher Information matrix $\mathcal{I}(\boldsymbol{\vartheta}) := -\mathbb{E}[\nabla_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}^\top}\ell(\boldsymbol{\vartheta})] = -\mathbf{D}^\top\overline{\mathbf{W}}\mathbf{D}$, where $\mathbf{W} = \overline{\mathbf{W}} + o_p(1)$ in the large sample approximation. It holds that $\overline{\mathbf{W}} := \text{diag}(\overline{\mathbf{W}}_1, \ldots, \overline{\mathbf{W}}_n, \mathbf{0}_p)$ because $\mathbf{K} = o_p(1)$. We next proceed, in analogy to GLMs, to the derivation of the corresponding penalized iteratively re-weighted least square (P-IRLS) algorithm (Green, 1984).

Assume that in the vicinity of the solution of (9) the corresponding step behaves like an unconstrained one, that $\mathbf{D}$ is of full rank p and $\overline{\mathbf{W}}$ is positive-definite throughout the parameter space. Then, from a quadratic approximation of $\ell_{\text{p}}(\boldsymbol{\vartheta}, \boldsymbol{\lambda}|\cdot)$ about $\boldsymbol{\vartheta}^{[\alpha+1]}$ we obtain, as unique

solution of the resulting non-singular $p \times p$ system of equations for $\boldsymbol{\vartheta}^{[\alpha+1]}$,

$$\begin{aligned} \boldsymbol{\vartheta}^{[\alpha+1]} &= \boldsymbol{\vartheta}^{[\alpha]} + (\mathcal{I}^{[\alpha]} + \mathbf{S}_{\boldsymbol{\lambda}}|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}})^{-1}(\mathbf{S}_{\boldsymbol{\lambda}}|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}}\boldsymbol{\vartheta}^{[\alpha]} - \boldsymbol{g}^{[\alpha]}) \\ \boldsymbol{\vartheta}^{*} &= (\mathbf{D}^\top\overline{\mathbf{W}}\mathbf{D} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{D}^\top\overline{\mathbf{W}}\mathbf{z}. \end{aligned}$$

For notational convenience, we have labelled $\boldsymbol{\vartheta}^* := \boldsymbol{\vartheta}^{[\alpha+1]}$ and ignored the superscript $[\alpha]$ in all the other quantities. Remarkably, these expressions involve arrays from the un-penalized log-likelihood, so that the only dependence on the smoothing parameters is through $\mathbf{S}_{\boldsymbol{\lambda}}$. So re-written, we observe that $\boldsymbol{\vartheta}^*$ is the solution of a Generalized Least Squares (GLS) normal equations problem, that is

$$\boldsymbol{\vartheta}^* = \arg\min_{\boldsymbol{t}} \left\|\overline{\mathbf{W}}^{1/2}(\mathbf{z} - \mathbf{D}\boldsymbol{t})\right\|^2 + \boldsymbol{t}^\top\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{t} \tag{10}$$

for any given value of $\boldsymbol{\lambda}$. In particular, $\overline{\mathbf{W}}^{1/2}$ comes from the spectral decomposition of $\overline{\mathbf{W}}$, whose computation is fostered by its construction as a block diagonal matrix. In other words, at each iteration the estimating algorithm solves a linear regression of $\mathbf{z}$ onto the columns of $\mathbf{D}$ with weight matrix $\overline{\mathbf{W}}$ and ridge penalisation $\boldsymbol{\vartheta}^\top\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\vartheta}$.

With this equivalence at hand, define now $\widehat{\boldsymbol{\mu}} := \overline{\mathbf{W}}^{1/2}\mathbf{D}\boldsymbol{\vartheta}^* = \mathbf{P}_{\boldsymbol{\lambda}}\overline{\mathbf{W}}^{1/2}\mathbf{z}$ to be the plug-in estimator of the mean of $\overline{\mathbf{W}}^{1/2}\mathbf{z}$ evaluated at (10), and let $\mathbf{P}_{\boldsymbol{\lambda}}$ be the influence matrix

$$\mathbf{P}_{\boldsymbol{\lambda}} = \overline{\mathbf{W}}^{1/2}\mathbf{D}(\mathbf{D}^\top\overline{\mathbf{W}}\mathbf{D} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{D}^\top\overline{\mathbf{W}}^{1/2}.$$

Hence we propose to select $\boldsymbol{\lambda}$ through the minimisation of the expected discrepancy between the true and the fitted curves:

$$\begin{aligned} \widetilde{n}^{-1}\mathbb{E}\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2 &= \widetilde{n}^{-1}\mathbb{E}\left\|\overline{\mathbf{W}}^{1/2}\mathbf{z} - \mathbf{P}_{\boldsymbol{\lambda}}\overline{\mathbf{W}}^{1/2}\mathbf{z} - \boldsymbol{\varepsilon}\right\|^2 \\ &= \widetilde{n}^{-1}\mathbb{E}\left[\left\|\overline{\mathbf{W}}^{1/2}(\mathbf{z} - \mathbf{D}\boldsymbol{\vartheta}^*)\right\|^2 + \|\boldsymbol{\varepsilon}\|^2 - 2\left\langle\overline{\mathbf{W}}^{1/2}\mathbf{z} - \mathbf{P}_{\boldsymbol{\lambda}}\overline{\mathbf{W}}^{1/2}\mathbf{z}; \boldsymbol{\varepsilon}\right\rangle\right] \\ &= \widetilde{n}^{-1}\mathbb{E}\left\|\overline{\mathbf{W}}^{1/2}(\mathbf{z} - \mathbf{D}\boldsymbol{\vartheta}^*)\right\|^2 - 1 + 2\widetilde{n}^{-1}\mathrm{tr}(\mathbf{P}_{\boldsymbol{\lambda}}), \end{aligned} \tag{11}$$

where $\widetilde{n} := 5n$. The last line is recovered by expanding the inner product $\langle\cdot\rangle$, and by constructing $\overline{\mathbf{W}}^{-1/2}\mathbf{u} =: \boldsymbol{\varepsilon} \sim (\mathbf{0}_{\widetilde{n}}, \mathbf{I}_{\widetilde{n}})$, the stochastic component of the GLS model leading to estimator (10). The trace of the influence matrix that appears in (11), computed by $\mathrm{tr}(\mathbf{P}_{\boldsymbol{\lambda}}) = \mathrm{tr}(\mathcal{I}_{\mathrm{p}}^{-1}\mathcal{I})$, defines the effective degrees of freedom (edf) of the model. They usually differ from the number of parametric model components because of the presence of the penalty matrix which can suppress some dimensions of the parameter space. Multiple smoothing parameter selection can then be performed via minimisation of (11), an estimator that is commonly termed UBRE and that reads as

$$\begin{aligned} \boldsymbol{\lambda}^{[\alpha+1]} &:= \arg\min_{\boldsymbol{\lambda}} \mathcal{V}_u(\boldsymbol{\lambda}) \\ &:= \left\|\overline{\mathbf{W}}^{1/2[\alpha+1]}(\mathbf{z}^{[\alpha+1]} - \mathbf{D}^{[\alpha+1]}\boldsymbol{\vartheta}^{[\alpha+1]}|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}})\right\|^2/\widetilde{n} - 1 + 2\mathrm{tr}(\mathbf{P}_{\boldsymbol{\lambda}})|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}}/\widetilde{n}. \end{aligned}$$

Alternative ways to select $\boldsymbol{\lambda}$ can be defined starting from the working linear model (10): the GCV, for example, is also left as an option in our routine. The corresponding criterion is given explicitly by Wood (2006).

As previously anticipated, a link between (11) and the log-likelihood AIC exists. In fact, upon approximating $-2\ell(\boldsymbol{\vartheta}^*)$ about $\boldsymbol{\vartheta}$, it can be shown that

$$-2\ell(\boldsymbol{\vartheta}^*) \approx -2\ell(\boldsymbol{\vartheta}) - \left\|\overline{\mathbf{W}}^{-1/2}\mathbf{u}\right\|^2 + \left\|\overline{\mathbf{W}}^{1/2}(\mathbf{z} - \mathbf{D}\boldsymbol{\vartheta}^*)\right\|^2.$$

Hence, by realising that the smoothing parameter vector enters the above expression only through $\boldsymbol{\vartheta}^*$, dropping all irrelevant terms yields

$$\mathcal{V}_u(\boldsymbol{\lambda}) \propto -2\ell(\boldsymbol{\vartheta}^*) + 2\mathrm{tr}(\mathbf{P}_{\boldsymbol{\lambda}}).$$

Algorithm 1 Computation of the MPLE within a Trust-region Optimisation Routine

Require: $\alpha \in (0, \texttt{iter.max})$; $d \in [0, 1/4]$; $\bar{\Delta} > 0$; $\kappa \geq 1$
  $\boldsymbol{\vartheta}^{[0]}$, $\boldsymbol{\lambda}^{[0]}$, $\boldsymbol{p}^{[0]}$, $\Delta^{[0]} \in (0, \bar{\Delta})$
  while $\alpha \leq \texttt{iter.max}$ or $\max \left| \boldsymbol{\vartheta}^{[\alpha+1]} - \boldsymbol{\vartheta}^{[\alpha]} \right| \geq 10^{-6}$ **do**
    $\boldsymbol{p}^{[\alpha+1]} \leftarrow \min_{\boldsymbol{p}} - \left[\ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}|\boldsymbol{\lambda}^{[\alpha]}) + \boldsymbol{p}^\top \nabla_{\boldsymbol{\vartheta}^{[\alpha]}} \ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}|\boldsymbol{\lambda}^{[\alpha]}) + \frac{1}{2} \boldsymbol{p}^\top \nabla_{\boldsymbol{\vartheta}^{[\alpha]}\boldsymbol{\vartheta}^{[\alpha]\top}} \ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}|\boldsymbol{\lambda}^{[\alpha]}) \boldsymbol{p} \right]$
      s.t. $\|\boldsymbol{p}\| \leq \Delta^{[\alpha]}$
    $\varrho^{[\alpha+1]} \leftarrow \left[\ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}|\boldsymbol{\lambda}^{[\alpha]}) - \ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]} + \boldsymbol{p}^{[\alpha+1]}|\boldsymbol{\lambda}^{[\alpha]}) \right] / \left[\tilde{\ell}_{\mathrm{p}}(\mathbf{0}|\boldsymbol{\lambda}^{[\alpha]}) - \tilde{\ell}_{\mathrm{p}}(\boldsymbol{p}^{[\alpha+1]}|\boldsymbol{\lambda}^{[\alpha]}) \right]$
    if $\varrho^{[\alpha+1]} < 1/4$ **then**
      $\Delta^{[\alpha+1]} \leftarrow 1/4 \Delta^{[\alpha]}$
    else
      if $\varrho^{[\alpha+1]} > 3/4$ and $\|\boldsymbol{p}^{[\alpha+1]}\| = \Delta^{[\alpha]}$ **then**
        $\Delta^{[\alpha+1]} \leftarrow \min(2\Delta^{[\alpha]}, \bar{\Delta})$
      else
        $\Delta^{[\alpha+1]} \leftarrow \Delta^{[\alpha]}$
      end if
    end if
    if $\varrho^{[\alpha+1]} > d$ **then**
      $\boldsymbol{\vartheta}^{[\alpha+1]} \leftarrow \boldsymbol{\vartheta}^{[\alpha]} + \boldsymbol{p}^{[\alpha+1]}$
    else
      $\boldsymbol{\vartheta}^{[\alpha+1]} \leftarrow \boldsymbol{\vartheta}^{[\alpha]}$
    end if
    $\boldsymbol{\lambda}^{[\alpha+1]} \leftarrow \min_{\boldsymbol{\lambda}} \left[\left\| \overline{\mathbf{W}}^{1/2[\alpha+1]} (\mathbf{z}^{[\alpha+1]} - \mathbf{D}^{[\alpha+1]} \boldsymbol{\vartheta}^{[\alpha+1]}|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}}) \right\|^2 / \widetilde{n} - 1 + 2\kappa \mathrm{tr}(\mathbf{P}_{\boldsymbol{\lambda}}|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}}) / \widetilde{n} \right]$
  end while

The steps described in these sections are made operative by adapting to the present context the outer iteration algorithm of O'Sullivan et al. (1986), which is detailed in Algorithm 1. In empirical analyses, however, the fitted tuning parameters may result in curves' estimates that are believed to be too wiggly by the researcher. If that is the case, the trace of the influence matrix can be inflated by a scaling parameter $\kappa \geq 1$ to obtain smoother functions. We refer the reader to Kim and Gu (2004) for more details on this point.

3.4 Further Results and Inference

At convergence of the optimisation algorithm, point-wise confidence intervals for the estimated non-parametric curve $\widehat{s}_{j,l_j}$ can be obtained from the distribution

$$\mathcal{N}(s_{j,l_j}(\mathrm{v}_{j,l_j,i}), \boldsymbol{b}_{j,l_j,i}^\top \mathbf{V}_{\boldsymbol{\vartheta},[j,l_j]} \boldsymbol{b}_{j,l_j,i}),$$

where $\mathbf{V}_{\boldsymbol{\vartheta},[j,l_j]}$ denotes the sub-matrix of $\mathbf{V}_{\boldsymbol{\vartheta}}$ corresponding to the parameters associated to the (j,l_j)-th smooth, and $\mathbf{V}_{\boldsymbol{\vartheta}} := -\mathcal{H}_{\mathrm{p}}^{-1}$ is the covariance matrix of the posterior distribution of $\boldsymbol{\vartheta}|\mathbf{w} \sim \mathcal{N}_p(\widehat{\boldsymbol{\vartheta}}, \mathbf{V}_{\boldsymbol{\vartheta}})$, with $\mathbf{w} := \mathbf{D}^\top \overline{\mathbf{W}} \mathbf{z}$. For the smooth functions included in the model $\mathbf{V}_{\boldsymbol{\vartheta}}$ is usually preferred to the more intuitive estimator $\mathbf{V}_{\widehat{\boldsymbol{\vartheta}}} := -\mathcal{H}_{\mathrm{p}}^{-1} \mathcal{H} \mathcal{H}_{\mathrm{p}}^{-1}$. In fact, as Marra and Wood (2012) showed in the context of GAMs, the former includes both a bias and a variance components in a frequentist sense, a feature that is not shared by $\mathbf{V}_{\widehat{\boldsymbol{\vartheta}}}$.

The construction of the posterior distribution above was firstly advocated by Wahba (1983) and Silverman (1985). They recognised that any penalised estimation framework has a natural counterpart in the explication of some prior believes about the likely features of the true model. In particular, the imposition of a conjugate normal prior for $\boldsymbol{\vartheta}$ assumes that smoother models are more probable than wiggly ones, whilst the same probability density is assigned to all models of

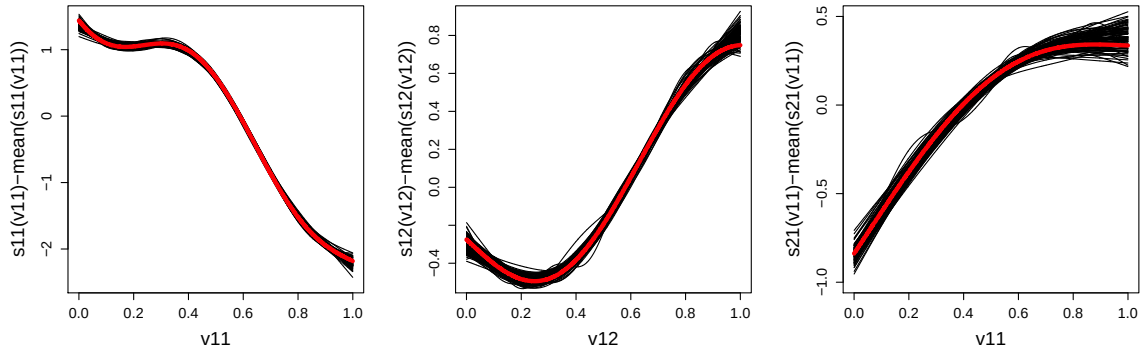

Figure 1: Estimated smooth curves obtained from 100 replicates of a Monte Carlo experiment comprising 10,000 simulated observations (true curves in red). Parameters' values were set close to the ones recovered in fitting the empirical illustration, in particular we have defined $\psi = -0.3$ and $\rho = 0.2$. The smooth components were represented using penalized thin plate regression splines with basis dimensions equal to 10 and penalties based on second-order derivatives. Results are plotted on the scale of the linear predictors. Please refer to the on line Supplementary Material for the exact definition of the DGP employed.

equal smoothness. Therefore, combining this reasoning with the normality of $\mathbf{w}$ (Wood, 2006), the stated result emerges.

Confidence intervals for non-linear functions of the parameter vector $\boldsymbol{\vartheta}$ can also be constructed using a convenient simulation scheme from the posterior distribution of $\boldsymbol{\vartheta}|\mathbf{w}$. We articulate the corresponding algorithm as follows. Let $T(\boldsymbol{\vartheta})$ be any function of the parameters, then

[1] draw N_{sim} vectors $\boldsymbol{\vartheta}^*_r, r = 1, \ldots, N_{sim}$, from $\mathcal{N}_p(\widehat{\boldsymbol{\vartheta}}, \mathbf{V}_{\boldsymbol{\vartheta}}(\widehat{\boldsymbol{\vartheta}}))$, where $\widehat{\boldsymbol{\vartheta}}$ is the MPLE;

[2] compute $T^*_r := T(\boldsymbol{\vartheta}^*_r)$ for every r, and define T^*_α to be the $[N_{sim}\alpha]$-th smallest value of the ordered sample $\{T^*_{(1)}, \ldots, T^*_{(N_{sim})}\}$, with $[a]$ denoting the integer part of $a \in \mathbb{R}$;

[3] obtain an approximate $(1-\alpha)\%$ confidence interval for $T(\widehat{\boldsymbol{\vartheta}})$ using $[T^*_{\alpha/2}, T^*_{1-\alpha/2}]$.

To gain insights into the effectiveness of the estimation approach, the results from a small Monte Carlo simulation study are presented in Figure 1. For the sake of conciseness, the exact definition of the Data Generating Process (DGP) is provided in the Supplementary Material. On average, the experiment shows that our method appears to be effective in recovering the true functions, although with a higher degree of uncertainty for the smooth in the simulated equation of Y_2 (last panel in the figure). This result is not unexpected. The recursive formulation of the model implies that the curves defining Y_1 enter the second equation directly through reduced-form system (2). Hence estimation of the corresponding parameters has to account also for this further source of uncertainty which stems from the first equation of the model. The same experiment has been repeated for $n = 3,000$ after the suggestion of one reviewer (see Supplementary Material). A similar pattern of Figure 1 is maintained when the sample size is reduced, although the uncertainty in recovering the curves is more evident in this case.

Some Asymptotic Considerations The large sample behaviour of the MPLE can be established under the relatively mild conditions of the consistency of the MLE. Following the arguments of Kauermann (2005), let us define

$$\boldsymbol{\vartheta}_0 := \arg\min_{\boldsymbol{\vartheta}} \mathrm{KL}(\mathcal{L}_t|\mathcal{L}_n) = \mathbb{E}[\ell_t - \ell_n(\boldsymbol{\vartheta})]$$

be the minimiser of the Kullback-Leibler discrepancy between the true structure that has generated the data and the employed model, and set the spline bases at a fixed high dimension.

This is a rather convenient assumption, but still of some relevance in applied research where the bases' dimension has to be fixed in order to achieve estimation. An existing drawback, however, is that the unknown smooth functions may not have an exact representation as linear combinations of the given bases at a finite dimension. Hence they may not be asymptotically recovered by their estimators as the sample size increases. Nonetheless, by using a number of bases rich enough to obtain a good representation of the unknown curves, it is possible to assume heuristically that the approximation bias is negligible compared to estimation variability (Kauermann, 2005).

Further let the following conditions hold: (i) $\nabla_{\boldsymbol{\vartheta}_0}\ell_n = O_p(n^{1/2})$, (ii) $\mathbb{E}[\nabla_{\boldsymbol{\vartheta}_0\boldsymbol{\vartheta}_0^\top}\ell_n] = O(n)$, (iii) $\nabla_{\boldsymbol{\vartheta}_0\boldsymbol{\vartheta}_0^\top}\ell_n - \mathbb{E}[\nabla_{\boldsymbol{\vartheta}_0\boldsymbol{\vartheta}_0^\top}\ell_n] = O_p(n^{1/2})$, and (iv) $\mathbf{S}_{\boldsymbol{\lambda}} = o(n^{1/2})$. Assumptions (i)-(iii) are the usual ones for the MLE consistency, whereas the last one is equivalent to consider $\lambda_{j,m_j} = o(n^{1/2})$ for any j. This comes from the very construction of the penalty matrix, and from the fact that every $\mathbf{S}_{j,m_j}$ is asymptotically bounded. Then the MPLE can be proved to satisfy

$$\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0 = \mathbf{F}^{-1}(\boldsymbol{\lambda})(\nabla_{\boldsymbol{\vartheta}_0}\ell(\boldsymbol{\vartheta}_0) - \mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\vartheta}_0)[1 + o_p(1)], \tag{12}$$

where $\mathbf{F}^{-1}(\boldsymbol{\lambda}) = (\mathbf{S}_{\boldsymbol{\lambda}} - \mathbb{E}[\nabla_{\boldsymbol{\vartheta}_0\boldsymbol{\vartheta}_0^\top}\ell(\boldsymbol{\vartheta}_0)])^{-1}$, and the leading stochastic component in (12) has asymptotic order $O_p(n^{-1/2})$ as $n \to \infty$. The proof of this result is given in the Appendix.

4 The Effect of Compulsory and Higher Education on Drinking Behaviour in the UK

Alcohol misuse has serious effects on global health and is commonly regarded as the third major risk factor for premature deaths and disabilities in the world (World Health Organization, 2007). It is also linked to a number of pathological conditions (e.g., coronary heart disease, stroke, liver disease and various cancers). The level of alcohol consumption in the United Kingdom has been recently reported by the OECD to be above the average of the EU countries (10.6 liters per capita against an average of 10.1 in 2012) and, despite its gradual decline in the EU from 1980, it has remained stable in the UK since then (OECD, 2014). In a report by the Public Health England, the HM Government acknowledged that as many as $21,485$ people died in 2012 from alcohol-related causes out of a total of around nine-million adults who drink at levels that pose some risk to their health (Public Health England, 2014). This comes with high costs for the society too. It has been estimated a total annual cost of alcohol-related harm of 21bn GBP, with an impact of 3.5bn GBP a year in costs related to alcohol for the National Health Service (NHS). The harmful use of alcohol compromises both individual and social development. The Crime Survey for England and Wales 2012-13, for example, showed that 49% of all violent crimes was connected to alcohol, with peaks involving 69% of stranger and 38% of domestic violences. In addition, problem drinking by parents is thought to contribute to the development of physical, psychological and behavioural problems in children.

In this study, we aim at applying the ideas discussed in the paper to investigate the effect of education on alcohol consumption in Great Britain. This is a non-trivial relationship since the level of education can act at different levels, and its overall effect is theoretically ambiguous. Recently, Huerta and Borgonovi (2010) surveyed and further elaborated on this aspect. On the one hand, more educated individuals are argued to have access to a wider spectrum of information relating to healthy behaviours, and usually acquire the necessary skills to process them and to act accordingly (Brunello et al., 2008, Goldman and Smith, 2005). Hence they may have a deeper knowledge about the risks connected to alcohol abuse (Kenkel, 1991). On the other hand, however, education shapes labour market opportunities and the social context in which people operate. As a result, better educated individuals face in general fewer financial constraints and may be exposed to working environments where drinking is acceptable if not even expected. Alongside with this lack of social stigma, an active social life and a high sense of

self-control may lead these people to have more frequent and possibly heavier drinking sessions than those of their less educated peers.

In addition to these conflicting directions in the sought relation, a number of other studies have also acknowledged the relevance of the time preference in predicting alcohol consumption (see O'Donoghue and Rabin, 2000, Fehr, 2002 and Delaney et al., 2008 just to name but a few). In particular, they indicate that people generally show a high rate of time preference with respect to their drinking behaviour, so that it is commonly perceived to be myopic. In other words, individuals tend to be more willing to put the well-being deriving from alcohol intake in the present rather than in the future, and this occurs at the expense of possible health-related problems. Education is also well understood to be associated with time preference. This point has been raised by Sander (1995), Bratti and Miranda (2010) and references therein in the context of smoking margins, and by Fuchs (1982) and van der Pol (2011) in a more general setting. Disentangling the true association between education and drinking behaviour requires therefore to account for this possible source of omitted variable bias. In the words of van der Pol (2011): "both education decisions and health decisions involve trade-offs of outcomes over time. Individuals' time preferences [...] will therefore influence how individuals make intertemporal choices such as whether or not to invest in education, whether to save or borrow and whether to engage in health affecting behaviours such as smoking, drinking and drug use" (p. 917). Finally, combining the evidence of low time preference for the choice of education and the aforementioned myopic attitude towards alcohol consumption, one would expect individuals' time preference to drive the two variables of interests in opposite directions which corresponds, in the current formulation of the model, to a negative correlation coefficient, $\rho < 0$.

4.1 Data and Empirical Analysis

We fit the simultaneous equation system model proposed in this paper to data from the 1970 British Cohort Study (BCS70), a longitudinal dataset of all children born in the Great Britain from the 5th to the 11th of April 1970, for a total of 17,198 babies surveyed. Information on the maximum educational level attained by the participants, as well as data on their geographical location and drinking behaviour were collected in the 29-year follow-up survey, whereas all the remaining variables are from the 10-year follow-up. This choice has been made primarily for data availability and the lower level of attrition experienced at these waves: after a first screening of the answers, we have a sample size of 7,115 respondents against the original 10,405 as from the merging of the two waves considered. Notice that item non-response in our main drinking variable, self-reported quantity of alcohol intake in the week prior to the interview, is very low (30), whereas a higher proportion of incomplete responses (2,090) were collected for the British Ability Scales (BAS). This is a battery of cognitive and achievement tests submitted to individuals and accounted in the 10-year follow-up.

The corresponding empirical bivariate densities of the dependent variables of interest, "highest education achieved" and "alcohol consumption", are given in Table 2. We note that the majority of respondents attended at most the O-levels, the compulsory lower secondary educational qualification in the UK, whilst only few people completed the A-levels without proceeding to any kind of Higher Education (HE). Concerning alcohol intake, around 39% of cohort members had alcoholic drinks in a week time at a level (in terms of units) of potential harm for their health. This threshold has been set according to the NHS recommendations of 2-3 units a day for women, and 3-4 for men. After having translated the different types of beverages into the corresponding alcohol units, we have distinguished usual drinkers between whoever intakes units within the suggested weekly limits ($\leq$ 14 u/w, level 3), "just" in the limit (14-21 u/w, level 4), and above them ($>$ 21 u/w, level 5). The values provided refer to women and the corresponding amounts for men can be computed analogously from the daily NHS recommendations. The remaining levels 1 and 2 comprise people who declared themselves to be only occasional/not drinkers at all, and light drinkers, respectively.

Highest Education	Alcohol Consumption					Marginals
	1	2	3	4	5	
Up to O-levels	1,191 (16.74%)	733 (10.30%)	1,125 (15.81%)	401 (5.64%)	1,285 (18.06%)	4,735 (66.55%)
A-levels	91 (1.28%)	71 (1.00%)	123 (1.73%)	47 (0.66%)	137 (1.93%)	469 (6.59%)
Higher Education	286 (4.02%)	184 (2.59%)	553 (7.77%)	218 (3.06%)	670 (9.42%)	1,911 (26.86%)
Marginals	1,568 (22.04%)	988 (13.89%)	1,801 (25.31%)	666 (9.36%)	2,092 (29.40%)	7,115 (100.00%)

Table 2: Empirical distribution of the observed categories for the response variables in the BCS70 29-year follow-up. In brackets we have reported the corresponding fraction of the sample size. Alcohol consumption is categorical in the original survey and is represented here with levels ranging from 1: "less often/only on special occasions (1,414); never nowadays (399); never had an alcoholic drink (192); don't know (4); not answered (16)" to 5: whoever drinks above the NHS recommended limits. Notice that level 1 includes also those individuals who declared themselves to drink at least once in a week, but no information about amount of alcohol consumed is reported (322).

Our model specification follows the one proposed in the literature by Bratti and Miranda (2009) and Huerta and Borgonovi (2010) in a similar context, and controls for some childhood circumstances that are commonly associated with alcohol abuse (Caldwell et al., 2008, Droomers et al., 2003, Hemmingsson et al., 1999, Poulton et al., 2002). In particular, we include variables referring to the parental presence in children's life and their interest in children's education, maternal weekly working hours, the highest parental social class, ethnicity and home tenure. The precise definition of these variables, along with their corresponding labels in the dataset, are given in the Supplementary Material for replicability purposes. We have excluded from the equation of the alcohol consumption the score obtained by the respondents in the BAS at the age of 10, as it is generally understood to affect the highest level of education attained by the cohort members. Nonetheless, it is also unlikely that the results of a test sat at an early age can be a *direct* predictor of the quantity of alcohol intake or drinking frequency at the age of 29, but through its effects on educational achievements. The same variable was also excluded by Bratti and Miranda (2009, 2010) in studying the effect of education on drinking frequency and smoking intensity in a similar bivariate framework. The system of equations, in `R` notation, is then:

$$\begin{aligned} \texttt{edu}_i^* \;\sim\; & \texttt{mum.not.pres}_i + \texttt{dad.not.pres}_i + \texttt{mum.edu}_i + \texttt{dad.edu}_i + \texttt{s.class}_i + \\ & \texttt{eth.child}_i + \texttt{mum.int.edu}_i + \texttt{dad.int.edu}_i + \texttt{sex.b}_i + \texttt{home}_i + \\ & s(\texttt{mum.wrk.hr}_i) + s(\texttt{BAS.tot}_i) \end{aligned}$$

$$\begin{aligned} \texttt{drk5}_i^* \;\sim\; & \texttt{edu}_i^* + \texttt{mum.not.pres}_i + \texttt{dad.not.pres}_i + \texttt{mum.edu}_i + \texttt{dad.edu}_i + \texttt{s.class}_i + \\ & \texttt{eth.child}_i + \texttt{mum.int.edu}_i + \texttt{dad.int.edu}_i + \texttt{sex.b}_i + \texttt{home}_i + \texttt{region}_i + \\ & s(\texttt{mum.wrk.hr}_i), \end{aligned}$$

where the categorical covariates are included in the above formulae as `as.factors(·)`.

Results and Interpretation The estimated parameters obtained by employing the above model are given in Table 3 for any discrete covariate, and in Figure 2 for the continuous predictors. For comparison purposes, the same specification has also been used for the fully parametric version of the proposed model, whose estimates are only reported for ψ and ρ as the main parameters of interest.

Although the raw estimates are not interpretable *per se*, a quick assessment of the converged log-likelihoods show that some gains are indeed achieved by employing a semi-parametric model

TRIANGULAR SEMI-PARAMETRIC BIVARIATE ORDERED PROBIT REGRESSION														
Variables	Highest Education Achieved							Alcohol Consumption						
	Covariates' levels													
	level 2	level 3	level 4	level 5	level 6	level 7	level 8	level 2	level 3	level 4	level 5	level 6	level 7	level 8
Mother not present	-.605	.112	–	–	–	–	–	.125	-.111	–	–	–	–	–
	(1.031)	(.098)	–	–	–	–	–	(.844)	(.078)	–	–	–	–	–
Father not present	.747	.093	–	–	–	–	–	-.187	.062	–	–	–	–	–
	(.819)	(.055)	–	–	–	–	–	(.678)	(.045)	–	–	–	–	–
Mother highest education	-.101	.175	.238	.482	.514	–	–	-.065	.025	-.015	-.004	.148	–	–
	(.069)	(.042)	(.081)	(.072)	(.107)	–	–	(.054)	(.037)	(.072)	(.066)	(.095)	–	–
Father highest education	.119	.127	.229	.151	.492	–	–	.102	.086	.021	.047	-.085	–	–
	(.066)	(.047)	(.057)	(.096)	(.058)	–	–	(.055)	(.039)	(.050)	(.085)	(.055)	–	–
Social class	-.112	-.125	-.365	-.108	-.314	-.581	–	-.042	.021	-.085	-.054	-.053	-.023	–
	(.118)	(.057)	(.075)	(.072)	(.058)	(.128)	–	(.095)	(.049)	(.062)	(.062)	(.050)	(.094)	–
Ethnicity	-.688	-.783	–	–	–	–	–	.523	.682	–	–	–	–	–
	(.059)	(.098)	–	–	–	–	–	(.473)	(.090)	–	–	–	–	–
Mother interested in child's education	-.050	-.029	-.212	–	–	–	–	.027	.008	-.109	–	–	–	–
	(.098)	(.059)	(.202)	–	–	–	–	(.078)	(.048)	(.127)	–	–	–	–
Father interested in child's education	-.101	-.113	-.180	–	–	–	–	.060	.050	.030	–	–	–	–
	(.063)	(.040)	(.166)	–	–	–	–	(.050)	(.032)	(.112)	–	–	–	–
Gender	-.089	–	–	–	–	–	–	.647	–	–	–	–	–	–
	(.032)	–	–	–	–	–	–	(.026)	–	–	–	–	–	–
Home tenure	-.028	-.406	-.102	–	–	–	–	.072	.160	.168	–	–	–	–
	(.190)	(.050)	(.044)	–	–	–	–	(.164)	(.044)	(.038)	–	–	–	–
Region of residence (levels 9-11 omitted)	–	–	–	–	–	–	–	-.079	.006	.021	-.459	-.080	.079	.032
	–	–	–	–	–	–	–	(.046)	(.054)	(.051)	(.290)	(.069)	(.043)	(.053)
	semi-parametric													parametric
ψ	0.300													0.304
	(0.023)													(0.022)
ρ	-0.217													-0.221
	(0.032)													(0.031)
log-likelihood	-15,192.81													-15,207.73
No. observations	7,115													7,115

Table 3: Estimated parameters for the categorical covariates included in the proposed triangular semi-parametric probit model; standard errors are reported in round brackets under the corresponding estimates. A comparison between the regression splines and the purely parametric models is included at the bottom.

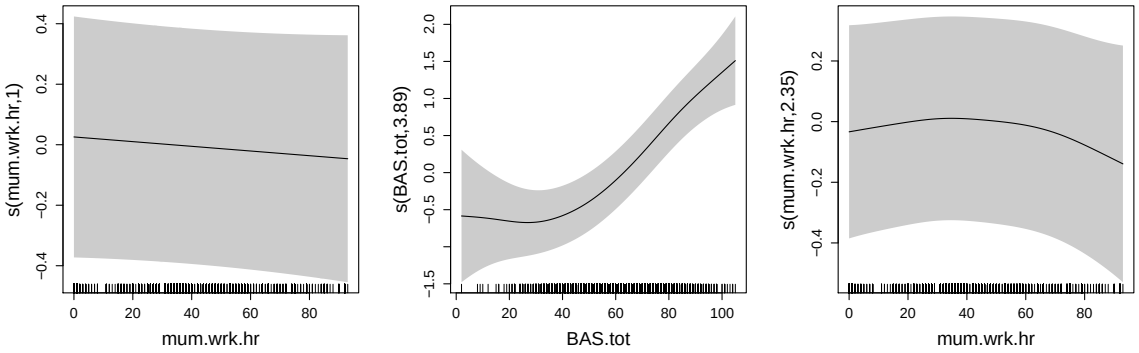

Figure 2: Estimated smooth functions and associated 95% point-wise confidence interval obtained by applying `SemiParCLM` to the BCS70 dataset. The first two curves correspond to the functions included in the equation for the educational achievements, while the last one to the model for the drinking frequency. The effective degrees of freedom are reported into brackets in the y-axis caption, with a value of one denoting the estimation of a straight line (as for the first curve). The actual covariate values are reported at the bottom of each graph through a jittered rug plot. The functions have been estimated using a low-rank penalized thin plate regression splines with basis dimensions equal to 10 and penalties based on second-order derivatives.

rather than just assuming linear covariate effects. At the same time, we note that the fitted values obtained from the two models are very close to each other. A possible explanation is that the effects of maternal weekly working hours are either estimated as a straight line, or are not important predictors for the responses. This conclusion is drawn from the observation that the zero line is entirely contained within the confidence intervals of the smooths. Hence, the mis-specification bias induced by a parametric functional form seems to be less amplified in this particular application. Quite interestingly, after having controlled for the possible source of omitted variables in the study, we find that childhood circumstances do not tend to be explanatory of the determination of both educational achievements and alcohol consumption of the cohort members. On the other hand, usual socio-demographic characteristics like parental education, social class, ethnicity and home tenure contribute to the explanation of children's highest level of education. This pattern is also confirmed by the estimated non-parametric curves, which appear to be uninformative in predicting the corresponding responses apart from the BAS values. To further check this, a shrinkage approach to variable selection in the spirit of Marra and Wood (2011) was performed (results are reported in the Supplementary Material for the sake of space). This method highlights that maternal working hours is not an influential predictor for the first equation, and hence it is safe to drop it from the current model specification. Final results remain, however, unchanged.

As previously anticipated, we can actually comment on the finding of a negative correlation among the two latent variables of the bivariate model. Specifically, if time preference is assumed to drive together the choice of education and the consumption of alcoholic beverages, the latter through its effect on undertaking healthier behaviours, then people who decide to invest in more schooling are also those less incline at recognising (or at considering) the consequences of alcohol abuse on their future health status. The estimated correlation coefficient is statistically different from zero, with a reported p-value for the null hypothesis $H_0 : \rho = 0$ of < 0.000. Therefore, the use of a simple univariate model which does not correct for the possible presence of omitted variables in the association of interest would have resulted in inconsistent estimates.

To give a better picture of the situation, we investigate the effects of education on people's weekly units of alcohol intake by looking at the predicted conditional probabilities (e.g. Greene and Hensher, 2010). Namely, we compute the probability of the average individual to consume a certain quantity of alcohol given his/her observed educational achievements. Formally, we

Highest Education	Alcohol Consumption				
			at least one drink per week		
	no/occasional	light	$<$ NHS limits	$\approx$ NHS limits	$>$ NHS limits
Up to O-levels	0.2320 (.1523; .3304)	0.1424 (.0432; .2230)	0.2550 (.2454; .2648)	0.0926 (.0833; .1016)	0.2780 (.2698; .2858)
A-levels	.2029 (.1300;.2957)	0.1354 (.0417; .2093)	0.2554 (.2457; .2651)	0.0967 (.0870; .1061)	0.3095 (.3010; .3181)
HE or equivalent	0.1876 (.1182;.2771)	0.1303 (.0405; .1998)	0.2530 (.2434; .2627)	0.0983 (.0885; .1079)	0.3307 (.3243; .3375)

Table 4: Average predicted conditional probabilities: each entry indicates the probability of a randomly drawn individual to have a certain weekly quantity of alcohol intake given his/her observed highest educational achievement. The 95% confidence intervals reported below the estimates are computed through simulation from the posterior distribution of $\boldsymbol{\vartheta}|\mathbf{w}$.

have

$$\widehat{\mathrm{PP}}_{k_2|k_1,i} := \frac{\mathbb{P}[y_{1,i}=k_1, y_{2,i}=k_2]}{\mathbb{P}[y_{1,i}=k_1]} = \left.\frac{\sum_{l,m\in\{0,1\}}(-1)^{l+m}\Phi_2(\eta_{1,k_1,i}(\boldsymbol{\vartheta}), \eta_{2,k_2,i}(\boldsymbol{\vartheta}); \boldsymbol{\Sigma})}{\Phi(\eta_{1,k_1,i}(\boldsymbol{\vartheta})) - \Phi(\eta_{1,k_1-1,i}(\boldsymbol{\vartheta}))}\right|_{\boldsymbol{\vartheta}=\widehat{\boldsymbol{\vartheta}}},$$

and the corresponding average effect is consequently $\widehat{\mathrm{APP}}_{k_2|k_1} = n^{-1}\sum_i \widehat{\mathrm{PP}}_{k_2|k_1,i}(\widehat{\boldsymbol{\vartheta}})$. The confidence intervals can be computed using the simulation approach detailed in Section 3.4. Table 4 finally reports the $\widehat{\mathrm{APP}}$ values for every combination of k_1 and k_2. In line with the theoretical arguments provided in the literature, we find that individuals with a HE qualification have a larger probability to intake weekly alcohol units above the NHS recommendations, and to drink more often than the lesser educated ones. The latter has been established by replacing the main drinking variable with a new one measuring the frequency of alcohol consumption during the week. Results are given in the Supplementary Material. In particular, the "degree effect" accounts for a 5.28% higher probability to drink at harmful levels compared to individuals who have (at most) completed the compulsory schooling. However, less education tends to be associated with a higher probability of being an occasional and/or a light alcohol consumer by 5.65% (2.05% if the individual got A-levels) with respect to people with a university degree.

As a note of caution, we warrant that the results obtained may not lend to an immediate generalisation to other contexts due to the very nature of the data analysed. In fact, although alcohol consumption is often regarded to vary with location and age, among other factors, cohort members were all born in the same week of 1970 in the UK, and their relevant drinking variables referring to the 29-year follow-up. Nonetheless, the reported association reveals, from a policy standpoint, that a raise in alcohol duties may not affect its (mis)consumption by the social group of educated young adults. In fact, because of the significant monetary wage returns of Higher Education (as documented, for example, by Blundell et al. 2000), this group may tend to be less price elastic, with a demand which is less responsive to a price change.

5 Concluding Remarks

In this paper we have introduced a bivariate triangular ordinal probit regression with semi-parametric covariate effect. Our model formulation has been recovered as an instance of a penalized Generalized Linear Model framework, so that estimation and inference have been conducted as a natural extension of GLMs. Semi-parametric modelling is of relevance in applications as it allows the researchers to achieve a higher degree of flexibility in empirical modelling. Hence it alleviates the bias arising from model mis-specification.

Following some relevant examples given in the literature (e.g. Sajaia 2008 and Buscha and Conte 2014), we have defined a prototypical recursive model with both ordinal polychotomous

responses collected in $\boldsymbol{Y} = (Y_1, Y_2)^\top$. This specification is usually employed in observational studies to account for the possible presence of unobserved confounding. Specifically, we have assumed that a response Y_2 of interest (alcohol consumption in the empirical illustration) is structurally dependent on a variable Y_1 (education achievements), and that a third factor affecting simultaneously the two is omitted from the analysis because not readily quantifiable (e.g., individual time preferences). In general, such an omission may induce a further source of association between Y_1 and Y_2 which is different from the relationship that the researcher is willing to investigate. This fact has been accounted for by estimating a correlation parameter ρ capturing the association implied by the confounder(s). Furthermore, we have identified the relationship between the elements in $\boldsymbol{Y}$ by including a variable which is independent of the time preference, does not affect the intake of alcohol units at the age of 29 (holding the educational achievements constant) and that is relevant in predicting the highest education of cohort members. These conditions define what is commonly regarded as an exclusion restriction in econometrics and epidemiology.

Incidentally, we have also illustrated how the triangular representation can be further qualified to recover other models nested in it, as well as the required modifications to be made in case one of the Y_j's is dichotomous. However, some directions remain to be explored. In particular, it could be of interest to investigate to what extent the representation proposed is applicable to different mixtures of discrete responses' types beyond the dichotomous/ordinal polychotomous one, or to extend the system of equations to encompass more than two dimensions. The further specification of the correlation coefficient as a function of some covariates is also of interest. This might help to investigate the role of unmeasured confounders in more depth, and to relate them to specific variables. The possible extension of the approach of Gertheiss and Tutz (2009) to the present context would also be useful to incorporate the implied monotonicity of the ordered covariates in our estimation algorithm. We will address these issues in future research.

Acknowledgments We are indebted to the Associate Editor and two anonymous reviewers whose many punctual comments have improved considerably the presentation of the article. We are grateful to the Centre for Longitudinal Studies (CLS), UCL Institute of Education for allowing us to use the BCS70 data and to the UK Data Service for making them available. However, neither CLS nor the UK Data Service bear any responsibility for the analysis or interpretation of these data.

Supplementary Material The on-line material attached includes the version of `SemiParCLM` used to derive the results presented, as well as the relevant codes to replicate the analysis. The dataset `BCS70.drk2` contains the variables employed in the model specification given in Section 4.1: original data are freely accessible upon registration at

`http://www.cls.ioe.ac.uk`.

The Supplementary Material file comprise two sections. The first describes the DGP used in the simulation studies and provides further evidence on the finite sample properties of the model. The second one includes some details on the empirical application proposed in this article.

A Proof of Result (12)

In addition to the previous assumptions (i)-(iv), we further assume the following: (v) for every $\vartheta^s \in \boldsymbol{\vartheta}$, $\partial^3/\partial\vartheta^{s3}(\ell_n(\boldsymbol{\vartheta}))$ exists and satisfies for every point $x \in \mathbb{R}$ and every parameter in the neighbourhood of ϑ_0^s: $|\partial^3/\partial\vartheta^{s3}(\ell_n(\boldsymbol{\vartheta}))| \leq M(x)$, with $\mathbb{E}[M(x)|\vartheta_0^s] < \infty$; and let $0 \leq \mathcal{I}(\vartheta_0^s) < \infty$.

Proof. We first set the notation. Let us denote by ϑ^j the j-th component of the parameter vector $\boldsymbol{\vartheta} = (\vartheta^1, \ldots, \vartheta^p)^\top$, and define $\ell_{\mathrm{p},j} := \partial\ell_\mathrm{p}/\partial\vartheta^j$ as the partial derivative of the penalized log-likelihood with respect to ϑ^j; higher order derivatives are denoted subsequently. Also, the "hat"

notation $\widehat{\ell}_{\mathrm{p}}$ stands for $\ell_{\mathrm{p}}(\widehat{\boldsymbol{\vartheta}})$, while the convention of omitting the listing of parameters is used wherever the relevant quantities are evaluated at the best coefficient $\boldsymbol{\vartheta}_0$, that is $\ell_{\mathrm{p}} := \ell_{\mathrm{p}}(\boldsymbol{\vartheta}_0)$.

Using the Einstein summation convention, we expand $\widehat{\ell}_{\mathrm{p},r}$ around $\ell_{\mathrm{p},r}$ using a second order Taylor approximation:

$$0 = \widehat{\ell}_{\mathrm{p},r} = \ell_{\mathrm{p},r} + \ell_{\mathrm{p},rs}(\widehat{\vartheta} - \vartheta_0)^s + \frac{1}{2}\ell_{\mathrm{p},rst}(\widehat{\vartheta} - \vartheta_0)^{st} + \cdots$$

with $(\widehat{\vartheta} - \vartheta_0)^s := \widehat{\vartheta}^s - \vartheta_0^s$ and $(\widehat{\vartheta} - \vartheta_0)^{st} = (\widehat{\vartheta} - \vartheta_0)^s(\widehat{\vartheta} - \vartheta_0)^t$. Solving the above equation for $\widehat{\vartheta} - \vartheta_0$, and denoting by superscripts the inverses of the respective quantities, we get (Barndorff-Nielsen and Cox, 1994):

$$(\widehat{\vartheta} - \vartheta_0)^r = -\ell_{\mathrm{p}}^{rs}\ell_{\mathrm{p},s} - \frac{1}{2}\ell_{\mathrm{p}}^{rtv}\ell_{\mathrm{p},u}\ell_{\mathrm{p},w} + \cdots \tag{13}$$

where $\ell_{\mathrm{p}}^{rtv} := \ell_{\mathrm{p}}^{rs}\ell_{\mathrm{p}}^{tu}\ell_{\mathrm{p}}^{vw}\ell_{\mathrm{p},stv}$, and ℓ_{p}^{rs} is the (r,s)-th element of the inverse observed (penalized) Fisher Information. Equation (13) can be simplified as follows (see, for example, Kauermann, 2005): $\ell_{\mathrm{p},rs} := f_{rs}(\lambda) + r_{rs}$, where $f_{rs}(\lambda) := f_{rs}(0) - s_\lambda^{rs}$ is the penalized expected Fisher Information contribution: $f_{rs}(0) := \mathbb{E}[\partial\ell/\partial\vartheta^r\partial\vartheta^s]$, and $r_{rs} := \ell_{rs} - f_{rs}(0)$.

Under assumptions (ii) and (iv) we find that $f_{rs}(\lambda)$ is of asymptotic order $O(n)$, and that $r_{rs} = O_p(n^{1/2})$ directly from (iii). We can then simplify the first term of (13) as

$$\begin{aligned} -\ell_{\mathrm{p}}^{rs} &= \mathbb{E}[\ell_{\mathrm{p},r}\ell_{\mathrm{p},s}]^{-1} + \mathbb{E}[\ell_{\mathrm{p},r}\ell_{\mathrm{p},t}]^{-1}\mathbb{E}[\ell_{\mathrm{p},s}\ell_{\mathrm{p},u}]^{-1}(\mathbb{E}[\ell_{\mathrm{p},t}\ell_{\mathrm{p},u}] + \ell_{p,tu}) \\ &= -f^{rs}(\lambda) + f^{rt}(\lambda)f^{su}(\lambda)(-f_{rs}(\lambda) + \ell_{p,tu}), \end{aligned}$$

that is $\ell_{\mathrm{p}}^{rs} = f^{rs}(\lambda) - f^{rt}(\lambda)f^{su}r_{tu}$; following now the argument of Kauermann et al. (2009) we have

$$\ell_{\mathrm{p}}^{rs} = f^{rs}(\lambda)[1 + O(n^{-1})O_p(n^{1/2})] = f^{rs}(\lambda)[1 + O_p(n^{-1/2})].$$

We next need to characterise the order of ℓ_{p}^{rtv}, which in turn depends on the one of $\ell_{\mathrm{p},stv}$. First note that $\ell_{\mathrm{p},stv} = \ell_{stv}$ from the very construction of the penalized likelihood estimator, so that we can safely apply (v), implying that we can bound in probability the third derivative of the log-likelihood. Then, by the strong law of large numbers, we have that, for almost every sequence of $\{x_1, \ldots, x_n\}$ and every $\boldsymbol{\vartheta} \in \Theta$,

$$|n^{-1}\ell_{stv}| \leq n^{-1}\sum_i M(x_i) \xrightarrow{as} \mathbb{E}[M(x)]$$

as $n \to \infty$, hence $n^{-1}\ell_{stv} = O_p(1)$. It is then implied $\ell_{stv} = O_p(n)$ and, after some tedious computations, $\ell_{\mathrm{p}}^{rtv} = f^{rs}(\lambda)f^{tu}(\lambda)f^{vw}(\lambda)O_p(n) = O_p(n^{-2})$ so that $\ell_{\mathrm{p}}^{rtv}\ell_{\mathrm{p},u}\ell_{\mathrm{p},w} = O_p(n^{-1})$ since $\ell_{\mathrm{p},u} = O_p(n^{1/2}) - o(n^{1/2})$. We also find that $\ell_{\mathrm{p}}^{rs}\ell_{\mathrm{p},s}$ has order $O_p(n^{-1/2}) + o(n^{-1/2})$, that is the second addendum in (13) becomes asymptotically negligible compared to $\ell_{\mathrm{p}}^{rs}\ell_{\mathrm{p},s}$. We can then write $(\widehat{\vartheta} - \vartheta_0)^r = -f^{rs}(\lambda)\ell_{\mathrm{p},s}[1 + o_p(1)]$, whose leading terms, in matrix notation, are $\mathbf{F}^{-1}(\boldsymbol{\lambda})(\nabla_{\boldsymbol{\vartheta}_0}\ell(\boldsymbol{\vartheta}_0) - \mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\vartheta}_0)$, from which the assertion follows.

The stochastic order of the above terms then stems from $f^{rs}(\lambda)\ell_{\mathrm{p},s} = O_p(n^{-1/2}) + o_p(n^{-1/2}) = O_p(n^{-1/2})$. ■

References

Aitchison, J. and Silvey, S. (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika*, 44(1/2):131–140.

Anderson, J. and Philips, P. (1981). Regression, discrimination and measurement models for ordered categorical variables. *Journal of the Royal Statistical Society, Series C*, 30(1):22–31.

Barndorff-Nielsen, O. and Cox, D. (1994). *Inference and Asymptotics*. Chapman & Hall, London, UK.

Blundell, R., Dearden, L., Goodman, D., and Reed, H. (2000). The returns to Higher Education in Britain: Evidence from a british cohort. *The Economic Journal*, 110(461):F82–F99.

Bratti, M. and Miranda, A. (2009). Selection-endogenous ordered probit and dynamic ordered probit models. *Proceedings of the United Kingdom Stata Users' Group Meetings 2009*.

Bratti, M. and Miranda, A. (2010). Non-pecuniary returns to Higher Education: The effect on smoking intensity in the UK. *Health Economics*, 19(8):906–920.

Brunello, G., Michaud, P., and Sanz-de Galdeano, A. (2008). The rise in obesity across the Atlantic: An economic perspective. *IZA Discussion Paper No. 3529*.

Buscha, F. and Conte, A. (2014). The impact of truancy on educational attainment during compulsory schooling: A bivariate ordered probit estimator with mixed effects. *The Manchester School*, 82(1):103–127.

Caldwell, T., Rodgers, B., Clark, C., Jefferis, B., Stansfeld, S., and Power, C. (2008). Lifecourse socioeconomic predictors of midlife drinking patterns, problems and abstention: Findings from the 1958 British Birth Cohort Study. *Drug and Alcohol Dependence*, 95(3):269–278.

Cox, D. and Wermuth, N. (2004). Causality: A statistical view. *International Statistical Review*, 72(3):285–305.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of Generalized Cross-Validation. *Numerische Mathematik*, 31(4):377–403.

Dale, J. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, 42(4):909–917.

Delaney, L., Harmon, C., and Wall, P. (2008). Behavioral economics and drinking behavior: Preliminary results from an Irish college study. *Economic Inquiry*, 46(1):269–272.

Droomers, M., Schrijvers, C., Casswell, S., and Mackenbach, J. (2003). Occupational level of the father and alcohol consumption during adolescence; patterns and predictors. *Journal of Epidemiology and Community Health*, 57(9):704–710.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.

Fehr, E. (2002). The economics of impatience. *Nature*, 415(6869):269–272.

Frosini, B. (2006). Causality and causal models: A conceptual perspective. *International Statistical Review*, 74(3):305–334.

Fuchs, V. (1982). *Economic Aspects of Health*, chapter Time Preference and Health: An Exploratory Study. University of Chicago Press, Chicago, IL.

Gertheiss, J. and Tutz, G. (2009). Penalized regression with ordinal predictors. *International Statistical Review*, 77(3):345–365.

Geyer, C. (2013). Trust regions. `http://cran.stat.ucla.edu/web/packages/trust/vignettes/trust.pdf`.

Goldman, D. and Smith, J. (2005). Socioeconomic differences in the adoption of new medical technologies. *American Economic Review*, 95(2):234–237.

Green, P. (1984). Iteratively Reweighted Least Squares for Maximum Likelihood estimation, and some robust and resistant alternatives (with discussion). *Journal of the Royal Statistical Society, Series B*, 46(2):149–192.

Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach*. Chapman & Hall, London, UK.

Greene, W. and Hensher, D. (2010). *Modeling Ordered Choices. A Primer*. Cambridge University Press, Cambridge, UK.

Haberman, S. (1980). Discussion of McCullagh (1980). *Journal of the Royal Statistical Society, Series B*, 42(2):136–137.

Heckman, J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46(4):931–959.

Hemmingsson, T., Lundberg, I., and Diderichsen, F. (1999). The roles of social class of origin, achieved social class and intergenerational social mobility in explaining social class inequalities in alcoholism among young men. *Social Science & Medicine*, 49(8):1051–1059.

Hillmann, J., Kneib, T., Koepcke, L., Paz, L., and Kretzberg, J. (2014). Bivariate cumulative probit model for the comparison of neuronal encoding hypotheses. *Biometrical Journal*, 56(1):23–43.

Huerta, M. and Borgonovi, F. (2010). Education, alcohol use and abuse among young adults in Britain. *Social Science & Medicine*, 71(1):143–151.

Kauermann, G. (2005). Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics & Data Analysis*, 49(1):169–186.

Kauermann, G., Krivobokova, T., and Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society, Series B*, 71(2):487–503.

Keane, M. (1992). A note on identification in the multinomial probit model. *Journal of Business & Economic Statistics*, 10(2):193–200.

Kenkel, D. (1991). Health behavior, health knowledge, and schooling. *Journal of Political Economy*, 99(2):287–305.

Kim, K. (1995). A bivariate cumulative probit regression model for ordered categorical data. *Statistics in Medicine*, 14(12):337–356.

Kim, Y. and Gu, C. (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *Journal of the Royal Statistical Society, Series B*, 66(2):337–356.

Klein, N. and Kneib, T. (2015). Simultaneous inference in structured additive conditional copula regression models: A unifying Bayesian approach. *Statistics and Computing (in press)*.

Klein, N., Kneib, T., Klasen, S., and Lang, L. (2015). Bayesian structured additive distributional regression for multivariate responses. *Journal of the Royal Statistical Society, Series C (in press)*.

Kosmidis, I. (2014). Improved estimation in Cumulative Link Models. *Journal of the Royal Statistical Society, Series B*, 76(1):169–196.

Marra, G. and Radice, R. (2011). Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. *The Canadian Journal of Statistics*, 39(2):259–279.

Marra, G. and Radice, R. (2013). A Penalized Likelihood estimation approach to semiparametric sample selection binary response modeling. *The Electronic Journal of Statistics*, 7:1432–1455.

Marra, G. and Wood, S. (2011). Practical variable selection for Generalized Additive Models. *Computational Statistics & Data Analysis*, 55(7):2372–2387.

Marra, G. and Wood, S. (2012). Coverage properties of confidence intervals for Generalized Additive Model components. *Scandinavian Journal of Statistics*, 39(1):53–74.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, 42(2):109–142.

McKelvey, R. and Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, 4(1):109–142.

Nelder, J. and Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, Series A*, 135(3):370–384.

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer, New York, NY.

O'Donoghue, T. and Rabin, M. (2000). The economics of immediate gratification. *Journal of Behavioural Decision Making*, 13(2):233–250.

OECD (2014). *Health at a Glance: Europe 2014*. OECD Publishing.

O'Sullivan, F., Yandell, B., and Raynor, W. (1986). Automatic smoothing of regression functions in Generalized Linear Models. *Journal of the American Statistical Association*, 81(393):96–103.

Peyhardi, J., Trottier, C., and Guédon, Y. (2014). A new specification of Generalized Linear Models for categorical data. *arXiv:1404.7331v2*.

Poulton, R., Caspi, A., Milne, B., Murray Thomson, W., Taylor, A., Sears, M., and Moffitt, T. (2002). Association between children's experience of socioeconomic disadvantage and adult health: A life-course study. *The Lancet*, 360(9346):1640–1645.

Public Health England (2014). *Alcohol Treatment in England 2013-2014*. Public Health England, London, UK.

Radice, R., Marra, G., and Wojtyś, M. (2015). Copula regression spline models for binary outcomes. *Statistics and Computing (in press)*.

Royston, P. and Altman, D. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (with discussion). *Journal of the Royal Statistical Society, Series C*, 43(3):429–467.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge, UK.

Sajaia, Z. (2008). Maximum Likelihood Estimation of a bivariate ordered probit model: Implementation and monte carlo simulations. *Unpublished manuscript*.

Sander, W. (1995). Schooling and quitting smoking. *The Review of Economics and Statistics*, 77(1):191–199.

Silverman, B. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society, Series B*, 47(1):1–52.

Snell, E. (1964). A scaling procedure for ordered categorical data. *Biometrics*, 20(3):592–607.

StataCorp (2015). *STATA: Data Analysis and Statistical Software: Release 13*.

UCL Institute of Education. Centre for Longitudinal Studies (2007). *Millennium Cohort Study: First Survey, 2001-2003 [computer file]*. UK Data Archive [distributor], Colchester, Essex, UK.

van der Pol, M. (2011). Health, education and time preference. *Health Economics*, 20(8):906–920.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society, Series B*, 45(1):133–150.

Wermuth, N. and Cox, D. (2008). Distortion of effects caused by indirect confounding. *Biometrika*, 98(1):481–493.

Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B*, 65(1):481–493.

Wood, S. (2004). Stable and efficient multiple smoothing parameter estimation for Generalized Additive Models. *Journal of the American Statistical Association*, 99(467):673–686.

Wood, S. (2006). *Generalized Additive Models. An Introduction With R*. Chapman & Hall/CRC, Boca Raton, FL.

World Health Organization (2007). *A60/14 Add.1, 60th World Health Assembly, Provisional Agenda Item 12.7*.

Yamamoto, T. and Shankar, V. (2004). Bivariate ordered-response probit model of driver's and passenger's injury severities in collisions with fixed objects. *Accident Analysis and Prevention*, 36(5):869–876.

Yee, T. and Wild, C. (1996). Vector Generalized Additive Models. *Journal of the Royal Statistical Society, Series B*, 58(3):481–493.

Zhang, Q. and Ip, E. (2012). Generalized Linear Model for partially ordered data. *Statistics in Medicine*, 31(1):56–68.